

Министерство науки и высшего образования Республики Казахстан  
Северо-Казахстанский университет  
им. М. Козыбаева

УДК: 004.032.26



На правах рукописи

**АУБАКИРОВА ГАИЛЯ ФАРИТОВНА**

**Разработка нейросетевой системы прогнозирования урожайности зерновых культур в Северо-Казахстанской области**

8D06101 Информатика, вычислительная техника и управление  
Диссертация на соискание ученой степени  
доктора философии (PhD)

Научный консультант:  
Доцент кафедры «Энергетика и  
радиоэлектроника» кандидат технических наук  
Герасимова Ю.В.

Зарубежный консультант:  
к. ф-м.н., заведующая кафедрой «Высшая  
математика» Сибирского  
государственного университета  
геосистем и технологий,  
г. Новосибирск, Россия;  
Григоренко О.В.

Республика Казахстан  
Петропавловск, 2026 г

## СОДЕРЖАНИЕ

<b>ВВЕДЕНИЕ</b>	4
<b>ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ ЗЕРНОВЫХ КУЛЬТУР</b>	8
1.1 Агрометеорологические факторы формирования урожайности зерновых культур в условиях Северо-Казахстанской области	8
1.2 Традиционные статистические методы прогнозирования урожайности	10
1.3 Методы машинного обучения в прогнозировании урожайности	14
<b>2 АНАЛИЗ МЕТОДОВ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ ЗЕРНОВЫХ КУЛЬТУР</b>	21
2.1 Математический аппарат прогнозирования урожайности (сельскохозяйственных) культур	24
2.2 Анализ существующих предикторов	28
2.2.1 Агрометеорологические параметры	29
2.2.2 Характеристики почвы	34
2.2.3 Биометрические параметры	35
2.3 Анализ существующих моделей урожайности	39
2.3.1 Модели урожайности на основе оценки природного потенциала территории	39
2.3.2 Агрометеорологические модели прогноза урожайности	42
<b>3 АППАРАТНАЯ РЕАЛИЗАЦИЯ СИСТЕМЫ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ</b>	49
3.1 Разработка структурной схемы	50
3.2 Проектирование электрической принципиальной схемы	52
3.3 Разработка печатной платы	56
<b>4 ИНТЕГРАЦИЯ ПОЛЕВЫХ ДАННЫХ В НЕЙРОСЕТЕВУЮ СИСТЕМУ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ</b>	75
4.1 Формирование, предварительная обработка и разведочный анализ локальных полевых данных	75
4.2 Проектирование и обучение локальной нейросетевой модели	83
<b>ЗАКЛЮЧЕНИЕ</b>	90
<b>СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ</b>	91
<b>ПРИЛОЖЕНИЕ А</b>	100
<b>ПРИЛОЖЕНИЕ В</b>	102

## ВВЕДЕНИЕ

**Актуальность работы.** Прогнозирование урожайности – важная, но сложная проблема, необходимая для устойчивой интенсификации и эффективного использования природных ресурсов [1, 2]. Заблаговременное и точное прогнозирование урожайности было и остаётся актуальной проблемой для любого государства, потому что эффективность длинной агропродовольственной цепочки зависит от точности прогноза. В этой цепочке участвуют фермеры, агрономы и политики и в своих действиях полагаются на прогнозы урожайности, данные экспертами [3-10].

Урожайность различных культур зависит от условий окружающей среды, управленческих действий и многих других специфических параметров [11]. Применяют различные подходы при прогнозировании урожайности, основные из них это: экспертные оценки (например, интервью и полевые исследования), статистические модели и модели, основанные на процессах [3]. Интервью с фермерами, как правило, предоставляют собой сугубо субъективные ожидания урожайности к концу сезона [12, 13]. Полевые исследования с обрезкой урожая дают объективную оценку урожайности до уборки урожая. Статистические модели применяют различные методы (регрессия, байесовские подходы, методы машинного обучения), для построения регрессионных зависимостей между различными статистическими данными, полученными при помощи дистанционных и метеорологических наблюдений [14, 15]. Одним из самых распространенных методов прогнозирования урожайности является статистическая модель на основе агрометеорологических данных. Данная модель относительно проста в разработке и использовании. Однако одним из главных недостатков этого метода является то, что многочисленные факторы окружающей среды (такие как температура воздуха, количество осадков), оказывающие наибольшее влияние на формирование урожайности пшеницы нелинейно, т.е. могут иметь большие отклонения от средних значений и поэтому необходимо отойти от традиционных методов в пользу более точных методов прогнозирования. Наиболее подходящей альтернативой являются модели, основанные на искусственных нейронных сетях.

Имитационные модели сельскохозяйственных культур производят не только урожайность в конце сезона, но также и распределение урожайности на основе генотипов сельскохозяйственных культур, состояния почвы, типичных методов управления и сезонной погоды. Эти данные получают на основе исторических прогнозов климата или погоды или путем усвоения информации, полученной с помощью дистанционного зондирования [16-18].

Преимуществом моделей с использованием нейронных сетей является высокая точность прогноза и возможность повышения урожайности. Алгоритмы построения и обучения нейронной сети базируются на функциях, которые определяют зависимость признаков и предикторов с выходными данными, в данном случае это урожайность. ИНС имеют некоторые явные преимущества,

перед традиционными моделями, так, они могут моделировать нелинейные отношения между несколькими источниками данных [19]. При этом их производительность обычно улучшается при большом количестве обучающих данных [20].

В разработку вопроса предвидения урожайности сельскохозяйственных культур внесли существенный вклад труды таких ученых как Н.Д. Кондратьев, В.М. Обухов, Н.А. Челинцев, Н. Четвериков, Е.М. Четыркин, С.С. Сергеев, С.Г. Струмилин, М.М. Юзбашев, В.Г. Сиротенко, В.А. Михельсон, А.М. Гатаулин, А.П. Зинченко, М.К. Каюмов, В.М. Пасов, Л.М. Перекальская, А.Н. Полевой, М.И. Розанов и другие.

В рамках данной диссертационной работы, выявление взаимосвязей агроклиматических, агрохимических, фитосанитарных и других значимых факторов, оказывающие существенное влияние на формирование урожая яровой пшеницы, позволят разработать методы мониторинга всех значимых факторов и разработать комплексную модель прогнозирования урожайности и валового сбора яровой пшеницы с использованием искусственной нейронной сети. Оценка ее достоверности, используя исторические данные агроклиматических, агрохимических, фитосанитарных и других значимых факторов и данные по урожайности конкретного поля позволят судить о преимуществах полученных прогностических моделей урожайности относительно существующих.

**Цель и задачи исследования.** Целью исследования является разработка системы прогнозирования урожайности с использованием ИНС, связывающую урожай пшеницы с количеством продуктивной влаги в почве, плодородием почвы, погодой и факторами наличия вредителей, болезней и сорняков для Северо-Казахстанской области, Республика Казахстан. Данная система может быть легко адаптирована для других культур и мест, а также может быть улучшена путем использования новых источников данных и более продвинутых функций.

Для достижения поставленных целей в диссертационной работе были определены и решены задачи:

- выявить агроклиматические, агрохимические, вегетационные и другие значимые факторы, оказывающие существенное влияние на формирование урожая яровой пшеницы;
- разработать методы мониторинга всех значимых факторов, оказывающие существенное влияние на формирование урожая яровой пшеницы;
- разработать комплексную модель прогнозирования урожайности и валового сбора яровой пшеницы с использованием искусственной нейронной сети;
- оценить достоверность комплексной модели прогнозирования урожайности, используя исторические данные агроклиматических, агрохимических, фитосанитарных и других значимых факторов и данные по урожайности конкретного поля.

**Объект исследования:** процесс прогнозирования урожайности зерновых культур в Северо-Казахстанской области.

**Предмет исследования:** методы и алгоритмы сбора и обработки данных, включая использование искусственной нейронной сети, для построения системы прогнозирования урожайности на основе различных факторов (агроклиматические, фитосанитарные данные, характеристики почвы).

**Научная новизна** исследования состоит в разработке концепции и методики проектирования нейросетевых систем прогнозирования урожайности зерновых культур, обеспечивающих комплексный учет агроклиматических, агрохимических, почвенных, фитосанитарных и оперативных полевых данных. Отличительной особенностью предложенного подхода является интеграция средств мониторинга, устройства сбора полевых данных и модели искусственной нейронной сети в единую систему прогнозирования, адаптированную к условиям Северо-Казахстанской области. Практическая реализация выполнена на примере яровой пшеницы.

**Научная новизна** заключается в следующих положениях и результатах:

– разработано автономное устройство собирающее и передающее агрометеорологические данные с исследуемого поля для дальнейшего построения прогнозной модели;

– разработана система сбора и обработки данных для прогноза урожайности зерновых культур в Северо-Казахстанской области на основе искусственной нейронной сети;

– выявлен базовый состав факторов, внесение которых в нейронную сеть повышает точность прогнозирования.

**Практическая ценность.**

1. Осуществлена аппаратная реализация системы прогнозирования урожайности зерновых культур, посредством внедрения в использование устройства сбора и передачи агрометеорологических данных.

2. Осуществлена программная реализация системы прогнозирования урожайности зерновых культур в среде MATLAB.

**Положения диссертации, выносимые на защиту:**

– система сбора, обработки и передачи агрометеорологических данных с поля;

– модель искусственной нейронной сети с набором данных, полученных с помощью устройства сбора и передачи агрометеорологических данных;

– алгоритм функционирования системы прогнозирования урожайности зерновых культур.

**Апробация работы.** Результат исследования система сбора, обработки и передачи агрометеорологических данных была внедрена на полях ТОО «Агрофирма «Мамлютское».

**Публикации.** Результаты диссертационного исследования опубликованы в 6 печатных научных работах, а именно, в 3 статьях, опубликованных в международных рецензируемых научных журналах, в 3 статьях, опубликованных в изданиях, включенных в Перечень научных изданий, рекомендуемых для

публикации основных результатов научной деятельности, утверждаемый уполномоченным органом, 1 инновационный патент РК.

**Личный вклад автора.** Основные результаты теоретических и экспериментальных исследований получены автором самостоятельно. В печатных работах, написанных в соавторстве соискателю принадлежит ведущая роль при анализе и обобщении полученных результатов.

**Структура диссертации.** Диссертация имеет классическую структуру: вводная часть, основная часть (четыре главы, в том числе выводы по каждой главе), заключение, список использованных источников. Работа изложена на 102 страницах компьютерного текста, включает 18 рисунка, 8 таблиц и 109 наименований библиографических источников.

# ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ ЗЕРНОВЫХ КУЛЬТУР

## 1.1 Агрометеорологические факторы формирования урожайности зерновых культур в условиях Северо-Казахстанской области

Северо-Казахстанская область относится к числу основных зернопроизводящих регионов Казахстана, где значительную долю в структуре растениеводства занимает возделывание яровой пшеницы. Формирование урожайности зерновых культур в данном регионе происходит в условиях резко континентального климата, для которого характерны значительная межгодовая изменчивость погодных условий, неравномерное распределение осадков и высокая зависимость продуктивности посевов от влагообеспеченности. В исследованиях по Северному Казахстану отмечается, что дефицит влаги является одним из ключевых факторов, ограничивающих повышение урожайности яровой мягкой пшеницы. В частности, А.Т. Babkenov и соавторы связывают необходимость селекции засухоустойчивых сортов с изменчивостью природно-климатических условий и высокой ролью влагообеспеченности в формировании продуктивности пшеницы в Северном Казахстане [1].

К основным агрометеорологическим факторам, влияющим на урожайность зерновых культур, относятся температурный режим, количество и распределение атмосферных осадков, запасы продуктивной влаги в почве, а также условия прохождения критических фаз роста и развития растений. Для яровой пшеницы важны не только суммарные значения температуры и осадков за сезон, но и их распределение по месяцам, поскольку недостаток влаги в отдельные периоды может совпадать с наиболее чувствительными фазами развития культуры. В исследовании R. Zholaman и соавторов, посвященном условиям Северо-Казахстанской равнины, показано, что продуктивность яровой мягкой пшеницы зависит от количества весенне-летних осадков и сроков сева [2].

Температурный режим оказывает влияние на скорость роста растений, продолжительность межфазных периодов и условия формирования генеративных органов. Для условий Северного Казахстана температурный фактор необходимо рассматривать во взаимосвязи с влагообеспеченностью: повышение температуры при недостатке осадков усиливает засушливость и может существенно снижать продуктивность посевов. Особое значение имеет период июнь–июль, когда у яровой пшеницы проходят критические этапы формирования продуктивного стеблестоя, колоса и зерна: дефицит влаги именно в этот период способен наиболее сильно ограничивать реализацию потенциальной урожайности. Это обосновывает использование в прогностических моделях помесячных агрометеорологических показателей, а не только среднегодовых или сезонных значений, что подтверждается исследованием М. Karatayev и соавторов, посвященным

мониторингу климата, засушливых условий и производства пшеницы в Казахстане [3].

Атмосферные осадки являются одним из наиболее значимых факторов формирования урожайности, однако их влияние определяется не только общим количеством, но и сроками выпадения. Недостаток осадков в активные фазы роста ограничивает формирование урожая, а избыточное увлажнение в отдельные периоды может способствовать развитию болезней и ухудшению условий созревания. В работе D.N. Nurpeisov и соавторов подчеркивается актуальность изучения изменчивости продуктивности яровой пшеницы в степной зоне Северного Казахстана в зависимости от характера выпадения осадков [4].

Почвенная влага является связующим показателем между погодными условиями и физиологическим состоянием растений. Даже при одинаковом количестве атмосферных осадков фактическая обеспеченность растений влагой может различаться в зависимости от предшествующих погодных условий, испаряемости, типа почвы и применяемых агротехнических мероприятий. Поэтому при прогнозировании урожайности важно учитывать не только температуру и осадки, но и влажность почвы на разных глубинах. В настоящей диссертационной работе это реализовано через включение в состав входных параметров нейросетевой модели трех показателей влажности почвы: на глубинах 0,1 м, 0,5 м и 1 м.

Дополнительную актуальность данной задаче придает изменение климата. В исследовании P. Romanovska и соавторов отмечается, что Северный Казахстан является важным регионом производства пшеницы, однако урожайность здесь существенно колеблется, а низкоурожайные годы повторяются достаточно часто. Авторы показывают влияние антропогенного изменения климата на производство пшеницы в Северном Казахстане и необходимость учета климатических рисков при анализе и прогнозировании урожайности [6].

Помимо агрометеорологических условий, на урожайность зерновых культур влияет фитосанитарное состояние посевов. Болезни, вредители и сорная растительность могут усиливать отрицательное воздействие неблагоприятных погодных условий, особенно в годы, благоприятные для развития патогенов и фитофагов. Для Северного Казахстана значимыми заболеваниями яровой пшеницы являются листовая ржавчина и септориоз. A. Morgounov, L. Rosseeva и M. Koushibayev отмечают, что Северный Казахстан и Западная Сибирь представляют собой крупную зону возделывания яровой пшеницы, где эпифитотии листовой ржавчины на восприимчивых сортах могут приводить к существенным потерям урожая [7].

Септориоз также рассматривается как важная болезнь пшеницы в Северном Казахстане. S.A. Vabkenova и соавторы исследовали видовой состав возбудителей септориозной пятнистости пшеницы в регионе и подчеркивают, что знание патогенного комплекса необходимо для создания и подбора устойчивых сортов [8]. Вредители яровой пшеницы также требуют учета при оценке условий

формирования урожая: Р.М. Уалиева и М.М. Каверина провели мониторинг фитофагов яровой пшеницы на Северо-Востоке Казахстана, что подтверждает актуальность включения фитосанитарных показателей в анализ продуктивности зерновых культур [9].

Комплексный учет перечисленных факторов обусловил формирование системы из 16 входных признаков нейросетевой модели, разработанной в настоящей диссертации: температура воздуха за май, июнь, июль и август; количество осадков за май, июнь, июль и август; показатели влажности почвы на глубинах 0,1 м, 0,5 м и 1 м; содержание азота и фосфора; а также индексы развития болезней, сорности и вредителей. Такой состав входных параметров позволяет учитывать как агрометеорологические, так и агробиологические условия формирования урожайности пшеницы в Северо-Казахстанской области.

## **1.2 Традиционные статистические методы прогнозирования урожайности**

Прогнозирование урожайности сельскохозяйственных культур является одной из ключевых задач аграрной науки, поскольку результаты такого прогноза используются для планирования производства, оценки продовольственной безопасности, управления агротехнологическими мероприятиями и принятия решений в условиях климатической неопределенности. До широкого распространения методов машинного обучения основой прогностических исследований выступали традиционные статистические методы: регрессионные модели, модели временных рядов, корреляционный анализ, трендовые модели, модели погодных индексов, а также панельные и пространственно-временные подходы. В систематическом обзоре В. Schaubерger, J. Jägermeуr и С. Gornott отмечается, что в задачах прогнозирования урожайности наиболее часто используются три группы входных данных: погодные данные, спутниковые данные и карты сельскохозяйственных культур [10].

Одним из наиболее распространенных традиционных подходов является регрессионное моделирование. В простейшем случае урожайность рассматривается как зависимая переменная, а температура воздуха, количество осадков, запасы влаги или другие агрометеорологические показатели — как независимые переменные. W. Shi, F. Tao и Z. Zhang выделяют три основные группы статистических моделей: модели временных рядов, поперечные модели и панельные модели [11]. Это деление важно для аграрных исследований, поскольку урожайность может анализироваться как во времени, так и по территориям.

Регрессионные модели широко применяются для оценки влияния температуры и осадков на урожайность. D.B. Lobell и M.B. Burke подчеркивают, что статистические модели являются распространенным инструментом для прогнозирования реакции урожайности на изменения средних значений температуры и осадков [12]. Преимущество таких моделей заключается в сравнительной простоте построения и возможности интерпретации

коэффициентов. Однако их точность во многом зависит от корректности выбора факторов, качества исходных данных и способности модели учитывать нелинейные реакции растений на погодные условия.

Развитие регрессионного подхода связано с использованием множественной линейной регрессии, где урожайность описывается через несколько независимых переменных одновременно. В исследовании Т. Conradt рассматривается применение множественных линейных регрессий для прогнозирования урожайности на основе погодных данных. Автор отмечает, что алгоритм ABSOLUT v1.2 проверяет большое количество комбинаций входных признаков для выбора наиболее информативных переменных, что показывает важную роль правильного отбора погодных показателей и временных интервалов их агрегации [13].

Для прогнозирования урожайности до уборки урожая также применяются регрессионные модели, основанные на погодной информации. Предуборочный прогноз представляет особую ценность для планирования логистики, принятия решений на зерновых рынках и распределения ресурсов в аграрном секторе. В.Н. Lee, Р. Kenkel и В.В. Brorsen разработали регрессионные модели для прогноза урожайности и качества пшеницы на уровне округов с использованием накопленных погодных данных за вегетационный период. Авторы показали, что детализация погодных показателей по неделям и фазам роста существенно повышает точность предуборочного прогноза [14].

Отдельное направление связано с анализом временных рядов, когда урожайность рассматривается как последовательность наблюдений за ряд лет, а прогноз строится на основе тренда, сезонности, автокорреляции и случайных колебаний. G. Bezabih и соавторы использовали анализ временных рядов для прогнозирования производства зерновых культур и показали применимость данного подхода при наличии длительных рядов наблюдений [15]. Для таких задач применяются авторегрессионные модели, ARIMA-подходы и метод экспоненциального сглаживания, однако их эффективность снижается, если урожайность резко изменяется под влиянием экстремальных погодных условий.

Статистические методы прогнозирования урожайности также включают модели погодных индексов, в которых агрометеорологические переменные агрегируются в комплексные показатели, отражающие условия роста в отдельные фазы вегетации. К.Н. Singh и соавторы применили LASSO-регрессию (Least Absolute Shrinkage and Selection Operator) в сочетании с погодными индексами для прогнозирования урожайности. Метод LASSO позволяет автоматически отбирать наиболее информативные предикторы путем их регуляризации, что особенно важно при большом числе входных переменных. В качестве входных переменных использовались максимальная и минимальная температура, осадки, утренняя и вечерняя относительная влажность в течение периода роста культуры; такой состав входных данных близок к подходу, реализованному в настоящей диссертации [16].

Традиционные статистические методы сохраняют значимость в задачах прогнозирования урожайности благодаря простоте реализации,

интерпретируемости результатов и возможности количественно оценивать влияние отдельных факторов. Линейные и множественные регрессионные модели позволяют установить направление и силу связи между урожайностью и агрометеорологическими показателями. Благодаря этому данные методы широко применяются как в агрометеорологических исследованиях, так и в практических задачах предварительной оценки урожайности [11].

Методы временных рядов, включая ARIMA и экспоненциальное сглаживание, позволяют учитывать многолетние тренды и закономерности изменения урожайности во времени. Модели класса ARIMA описывают зависимость текущего значения показателя от его прошлых значений и случайных компонентов, что делает их применимыми для задач долгосрочного прогнозирования динамики урожайности в регионе. Однако для оперативного прогноза в течение вегетационного периода применение моделей временных рядов может быть ограничено: без включения внешних предикторов — температуры, осадков или показателей влагообеспеченности — такие модели не способны учитывать специфику текущего сезона и реагировать на аномальные погодные условия.

Для условий Казахстана особое значение имеют статистические модели, основанные на погодных и урожайных данных. P. Romanovska, B. Schauburger и C. Gornott разработали статистическую модель прогнозирования урожайности пшеницы на уровне областей Казахстана с использованием общедоступных данных и подчеркивают, что такая модель требует относительно небольших вычислительных ресурсов и может быть воспроизведена на основе доступных данных [5].

Вместе с тем традиционные статистические методы имеют ряд ограничений. Линейные модели не всегда способны описывать сложные нелинейные связи между урожайностью и факторами среды: повышение температуры может положительно влиять на развитие растений до определенного порога, но при дальнейшем росте эффект становится отрицательным. Ограничением моделей временных рядов является их зависимость от исторической динамики — они хорошо описывают устойчивые тренды, но хуже учитывают резкие изменения. Как отмечают W. Shi, F. Tao и Z. Zhang, при использовании статистических моделей возникают методические сложности, связанные с выбором пространственного масштаба и структуры модели [11].

Таким образом, традиционные статистические методы прогнозирования урожайности являются важной методологической основой для аграрных исследований. Однако их ограниченная способность учитывать нелинейные зависимости, взаимодействие факторов и сложную пространственно-временную изменчивость создает предпосылки для применения более гибких методов моделирования, включая методы машинного обучения и искусственные нейронные сети. В рамках настоящего диссертационного исследования традиционные статистические подходы рассматриваются как теоретическая и методическая база,

на основе которой обосновывается переход к нейросетевому прогнозированию урожайности зерновых культур в условиях Северо-Казахстанской области.

Анализ научной литературы позволяет систематизировать существующие подходы к прогнозированию урожайности зерновых культур в единую классификационную схему. Все рассматриваемые методы могут быть разделены на две основные группы: традиционные статистические методы, включающие регрессионный анализ, модели временных рядов и корреляционные подходы, а также методы машинного обучения, в составе которых выделяются классические алгоритмы, нейронные сети и гибридные модели. Данная классификация представлена на рисунке 1.1.



Рисунок 1.1 – Классификация методов прогнозирования урожайности зерновых культур

Как следует из представленной классификации, обе группы методов имеют свою область применения и набор преимуществ. Традиционные статистические методы обеспечивают интерпретируемость результатов и применимы при ограниченных данных, однако не всегда способны описывать сложные нелинейные зависимости. Методы машинного обучения, и в особенности нейронные сети, обладают значительным потенциалом для задач прогнозирования урожайности в условиях многомерных нелинейных взаимосвязей между агрометеорологическими, почвенными и агробиологическими факторами. Именно эти свойства обусловили выбор нейросетевого подхода в качестве основного инструмента прогнозирования в настоящей диссертации, что подробнее рассматривается в разделе 1.3.

### 1.3 Методы машинного обучения в прогнозировании урожайности

В связи с выявленными ограничениями традиционных статистических методов – прежде всего их недостаточной способностью описывать нелинейные зависимости и взаимодействия между факторами – в современных исследованиях все активнее применяются методы машинного обучения. В отличие от регрессионных и авторегрессионных моделей, где форма зависимости задается заранее, методы машинного обучения способны автоматически находить закономерности в многомерных данных, включая нелинейные связи и

взаимодействия между признаками. В систематическом обзоре T. van Klompenburg, A. Kassahun и С. Catal отмечается, что в исследованиях по прогнозированию урожайности наиболее часто используются признаки, связанные с температурой, осадками, почвенными условиями и агротехнологическими параметрами [17].

В аграрной сфере машинное обучение используется для прогнозирования урожайности, мониторинга состояния растений, выявления болезней, классификации сельскохозяйственных культур, анализа почвенных и метеорологических данных, а также для задач точного земледелия. K.G. Liakos и соавторы в своем обзоре методов машинного обучения в сельском хозяйстве показали, что данные методы позволяют обрабатывать многомерные данные от сенсоров, спутниковых систем и других источников агромониторинга, что делает их важным инструментом цифровизации отрасли [18]. Авторы отмечают, что применение МО в агрономии значительно расширилось благодаря росту доступности данных и вычислительных мощностей, что позволило переходить от простых линейных моделей к более сложным нелинейным алгоритмам.

Одним из методов машинного обучения, применяемых для прогнозирования урожайности, является метод опорных векторов, включая его регрессионную форму (Support Vector Regression, SVR). Данный подход позволяет строить как линейные, так и нелинейные модели зависимости за счет использования ядерных функций. Его преимуществом является способность работать при сравнительно небольшом объеме данных и контролировать сложность модели. Однако эффективность метода существенно зависит от выбора ядра и параметров регуляризации. Применение SVR в задачах прогнозирования урожайности отмечается в систематических обзорах как один из распространенных базовых алгоритмов [17, 18].

Другую группу методов составляют деревья решений и ансамблевые алгоритмы, построенные на их основе. Дерево решений представляет собой модель, в которой прогноз формируется путем последовательного разбиения данных по значениям входных признаков. Преимуществом данного подхода является наглядность и простота интерпретации. Однако отдельное дерево решений может быть неустойчивым к изменению исходных данных и склонным к переобучению, что ограничивает его применение в прогнозировании урожайности как самостоятельного метода [17].

Для повышения устойчивости моделей применяются ансамблевые методы. Одним из наиболее известных алгоритмов является Random Forest, предложенный L. Breiman. Данный метод строит совокупность деревьев решений, каждое из которых обучается на случайной подвыборке данных и признаков, а итоговый прогноз формируется путем усреднения результатов [19]. J.H. Jeong и соавторы применили Random Forest для прогнозирования урожайности пшеницы, кукурузы и картофеля на глобальном и региональном уровнях и показали перспективность данного метода при анализе климатических и биофизических факторов [20].

К ансамблевым методам также относятся алгоритмы градиентного бустинга (XGBoost, LightGBM, CatBoost). В отличие от Random Forest, в алгоритмах бустинга каждое следующее дерево обучается с учетом ошибок предыдущих моделей, что часто обеспечивает высокую точность. А. Uribeetxebarria, A. Castellón и A. Aizpurua использовали алгоритм CatBoost для прогнозирования урожайности пшеницы на основе данных Sentinel-1 и Sentinel-2 [21]. Преимуществом бустинга является высокая прогностическая способность, однако такие модели требуют тщательной настройки гиперпараметров и строгого контроля переобучения.

Особое место среди методов машинного обучения занимают искусственные нейронные сети. Их преимущество заключается в способности моделировать сложные нелинейные зависимости без необходимости заранее задавать вид математической функции. Теоретической основой применения многослойных нейронных сетей является положение о том, что feedforward-сети при наличии достаточного числа скрытых нейронов обладают свойствами универсальных аппроксиматоров [22]. Для прогнозирования урожайности это особенно важно, поскольку продуктивность зерновых культур определяется совокупным действием температуры, осадков, влажности почвы, питательных веществ и фитосанитарных факторов.

Практическая эффективность нейронных сетей в задачах прогнозирования урожайности подтверждается современными исследованиями. S. Khaki и L. Wang применили глубокие нейронные сети для прогнозирования урожайности и показали способность таких моделей выявлять сложные зависимости между генотипическими, средовыми и управленческими факторами [23]. В дальнейшем S. Khaki, L. Wang и S.V. Archontoulis предложили CNN-RNN-фреймворк для прогнозирования урожайности, учитывающий временную динамику факторов среды [24].

Перспективным направлением является объединение методов машинного обучения с имитационными моделями роста сельскохозяйственных культур. M. Shahhosseini и соавторы показали, что сочетание crop modeling и machine learning может улучшать прогноз урожайности по сравнению с использованием отдельных подходов [25]. G. Lischeid и соавторы подчеркивают, что машинное обучение является мощным инструментом моделирования, однако не заменяет научного понимания процессов формирования урожая [26].

Для Казахстана применение методов машинного обучения особенно актуально, поскольку значительная часть зернового производства сосредоточена в регионах с высокой зависимостью от погодных условий. M. Sadenova и соавторы разработали подход к прогнозированию урожайности зерновых, зернобобовых, масличных и кормовых культур в Казахстане с использованием методов машинного обучения, нейронных сетей и данных дистанционного зондирования [27]. A. Mimenbayeva и соавторы также применяют методы машинного обучения для анализа и прогнозирования урожайности сельскохозяйственных культур с учетом погодных параметров [28].

Для систематизации рассмотренных подходов в таблице 1.3 приведена сравнительная характеристика основных методов машинного обучения, применяемых для прогнозирования урожайности сельскохозяйственных культур. Сравнение выполнено по четырем параметрам: основному принципу работы метода, его преимуществам, ограничениям применения и применимости к условиям задачи, решаемой в настоящей диссертации, — прогнозирования урожайности зерновых культур по агрометеорологическим, почвенным и фитосанитарным данным в разрезе 13 районов Северо-Казахстанской области за период 2007–2020 гг.

Таблица 1.3 – Сравнительная характеристика методов машинного обучения, применяемых для прогнозирования урожайности

Метод	Основной принцип	Преимущества	Ограничения
Метод опорных векторов / SVR	Построение регрессионной зависимости с использованием опорных объектов и ядерных функций	Позволяет описывать нелинейные зависимости; применим при ограниченном объеме данных; хорошо контролирует сложность модели	Требует тщательного выбора ядра и параметров; интерпретация результатов затруднена; при большом числе признаков настройка усложняется
Деревья решений	Последовательное разбиение данных по значениям входных признаков с формированием правил прогноза	Простая логика построения; высокая интерпретируемость; возможность выявлять пороговые значения факторов	Склонность к переобучению; нестабильность при изменении исходных данных; ограниченная обобщающая способность отдельного дерева
Random Forest	Ансамбль деревьев решений, обучаемых на случайных подвыборках данных и признаков	Устойчивость к шуму; способность учитывать нелинейные связи; возможность оценки значимости признаков	Более сложная интерпретация по сравнению с линейной регрессией; модель не всегда ясно показывает характер влияния каждого фактора
Градиентный бустинг / XGBoost, LightGBM, CatBoost	Последовательное построение моделей, каждая из которых исправляет ошибки предыдущей	Высокая точность прогноза; эффективная работа со сложными зависимостями; возможность обработки разнородных данных	Чувствительность к настройке гиперпараметров; риск переобучения при малой выборке; требуется контроль сложности модели
Нейронные сети / Многослойный перцептрон	Аппроксимация нелинейных зависимостей между входными признаками и	Способность моделировать сложные нелинейные зависимости; гибкость архитектуры; применимость к	Риск переобучения при малой выборке; необходимость выбора архитектуры и

	урожаем с помощью слоев нейронов и функций активации	многомерным агрометеорологическим данным	алгоритма обучения; менее наглядная интерпретация по сравнению с регрессией
Глубокие нейронные сети / CNN, RNN, CNN-RNN	Использование многослойных архитектур для извлечения пространственных, временных или комбинированных признаков из данных	Высокая выразительность при обработке временных рядов и пространственных данных; автоматическое извлечение признаков	Требуют больших объемов данных и вычислительных ресурсов; склонны к переобучению при малых выборках
Гибридные модели	Объединение машинного обучения с имитационными моделями роста культур или статистическими подходами	Сочетание агрономических знаний с гибкостью ML; повышение физической обоснованности и интерпретируемости прогноза	Сложность реализации; требуют специальных знаний в области агрономического моделирования и ML; высокие требования к данным

Как видно из таблицы 1.3, каждый метод машинного обучения имеет собственные преимущества и ограничения, которые необходимо учитывать при выборе инструмента для конкретной прикладной задачи. Методы опорных векторов и ансамблевые алгоритмы (RF, XGBoost) обладают высокой прогностической способностью и относительной устойчивостью к переобучению, однако их интерпретируемость ниже, чем у регрессионных моделей. Глубокие нейронные сети (CNN, RNN, LSTM) и гибридные модели демонстрируют наибольший потенциал при обработке больших и разнородных массивов данных, однако при ограниченной выборке риск переобучения существенно возрастает. Многослойный перцептрон (MLP) занимает промежуточное положение: он способен аппроксимировать сложные нелинейные зависимости и при этом эффективно обучается на выборках умеренного объема, особенно при использовании алгоритма Левенберга-Марквардта.

В связи с этим для настоящего диссертационного исследования наиболее обоснованным является применение многослойного перцептрона с алгоритмом Левенберга-Марквардта, который обеспечивает эффективное обучение нейронных сетей при нелинейных зависимостях и относительно небольших объемах данных. Алгоритм Левенберга-Марквардта сочетает преимущества метода градиентного спуска и метода Гаусса-Ньютона, что обеспечивает быструю сходимость и устойчивость обучения нейронной сети [29, 30].

Следовательно, выбранная в диссертационной работе нейросетевая модель является методологически обоснованным решением для прогнозирования урожайности зерновых культур в условиях Северо-Казахстанской области. Объем исходной базы данных составляет 182 наблюдения по 13 районам за 2007–2020 гг.

Модель достаточно гибкая для описания нелинейных агрометеорологических, почвенных и фитосанитарных зависимостей, а применение алгоритма Левенберга-Марквардта в сочетании с ансамблевой кросс-валидацией обеспечивает высокую обобщающую способность при ограниченном объеме данных. Это позволяет рассматривать многослойный перцептрон как базовую модель нейросетевой системы прогнозирования урожайности, разработанной в настоящем исследовании.

### **Выводы по главе 1**

Проведенный в первой главе анализ теоретических основ прогнозирования урожайности зерновых культур позволяет сформулировать следующие выводы.

Формирование урожайности яровой пшеницы в условиях Северо-Казахстанской области определяется совокупностью агрометеорологических, почвенных и фитосанитарных факторов. Ключевую роль играют температурный режим и количество атмосферных осадков в критические фазы вегетации (май–август), запасы продуктивной влаги в почве на различных глубинах, содержание питательных элементов, а также фитосанитарное состояние посевов. Высокая межгодовая изменчивость погодных условий и частая повторяемость засушливых лет обуславливают значительные колебания урожайности, что существенно затрудняет ее прогнозирование на основе простых линейных моделей. Данные обстоятельства обусловили выбор состава входных признаков нейросетевой модели: 16 параметров, охватывающих агрометеорологические, почвенные и агробиологические условия формирования урожая.

Анализ традиционных статистических методов показал, что регрессионные модели, модели временных рядов и корреляционный анализ обеспечивают интерпретируемость результатов и применимы при ограниченных объемах данных. Вместе с тем их ограниченная способность описывать нелинейные взаимодействия между факторами среды снижает точность прогноза в условиях изменчивого климата Северного Казахстана, где связи между урожайностью и агрометеорологическими показателями носят выраженный нелинейный характер.

Обзор методов машинного обучения показал, что нейронные сети, в особенности многослойный перцептрон, обладают доказанной способностью моделировать сложные нелинейные зависимости между урожайностью и многомерными входными данными. Применение алгоритма Левенберга-Марквардта обеспечивает устойчивое и быстрое обучение нейронной сети при относительно небольшом объеме выборки. Данный подход подтвержден как в мировой научной литературе, так и в исследованиях, проводившихся непосредственно в Казахстане. Таким образом, использование многослойного перцептрона с алгоритмом Левенберга–Марквардта в качестве ключевого метода прогнозирования урожайности зерновых культур в условиях Северо-Казахстанской области является теоретически обоснованным и методологически оправданным решением.

## 2 МОДЕЛЬ НЕЙРОСЕТЕВОГО ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ ПШЕНИЦЫ

### 2.1 Формирование исходного набора данных и системы входных признаков

В состав Северо-Казахстанской области входит 13 районов: Кызылжарский, М.Жумабаева и Аккайынский, Г.Мусрепова, Айыртауский, Есильский, Мамлютский, Тайыншинский, Тимирязевский, Жамбылский, Акжарский, Шалакынский и Уалихановский. Для разработки ИНС использовались данные вышеуказанных 13 районов Северо-Казахстанской области (рисунок 2.1).



Рисунок 2.1 – Карта районов Северо-Казахстанской области, включенных в исследование

Урожайность пшеницы варьируется внутри и между районами в зависимости от факторов окружающей среды, плодородия и влажности почвы, наличия болезней растений, вредителей и сорняков. Функционально это выражается как:

$$Y=f(AF, WF, SF), \quad (2.1)$$

$Y$  – урожайность пшеницы (ц/га);

$AF$  – агрохимические факторы;

$WF$  – метеорологические факторы;

$SF$  – фитосанитарные факторы.

В ходе исследования были рассмотрены различные факторы, потенциально влияющие на формирование урожайности пшеницы. Для построения искусственной нейронной сети были выбраны 16 входных признаков, характеризующих условия формирования урожая в разрезе каждого наблюдения

типа «район–год» (таблица 2.1). Входные данные были распределены по трем группам факторов: агрохимические, метеорологические и фитосанитарные. Такое деление позволяет учитывать комплексное влияние влагообеспеченности, питательного режима почвы, погодных условий и биологических ограничений на урожайность пшеницы.

Таблица 2.1 – Слои информации, использованные для построения ИНС

Тип	Входной слой			Выходной слой
	Фактор	Размерность	Обозначение	
Агрохимические факторы ( <i>AF</i> )	запас продуктивной влаги 0–10 см	мм	X <sub>1</sub>	урожайность ( <i>y</i> )
	запас продуктивной влаги 0–50 см	мм	X <sub>2</sub>	
	запас продуктивной влаги 0–100 см	мм	X <sub>3</sub>	
	Содержание азота	мг/кг	X <sub>4</sub>	
	Содержание фосфора	мг/кг	X <sub>5</sub>	
Метеорологические факторы ( <i>WF</i> )	Количество осадков за май	мм	X <sub>6</sub>	
	Количество осадков за июнь	мм	X <sub>7</sub>	
	Количество осадков за июль	мм	X <sub>8</sub>	
	Количество осадков за август	мм	X <sub>9</sub>	
	ср. температура воздуха за май	°С	X <sub>10</sub>	
	ср. температура воздуха за июнь	°С	X <sub>11</sub>	
	ср. температура воздуха за июль	°С	X <sub>12</sub>	
	ср. температура воздуха за август	°С	X <sub>13</sub>	
Фитосанитарные факторы ( <i>SF</i> )	болезни	у.е.	X <sub>14</sub>	
	сорность	у.е.	X <sub>15</sub>	
	вредители	у.е.	X <sub>16</sub>	

К группе *AF* в настоящем исследовании отнесены агрохимические показатели, характеризующие состояние почвенной среды. В состав данной группы включены запасы продуктивной влаги в слоях 0–10 см, 0–50 см и 0–100 см, а также содержание азота и фосфора. Объединение указанных параметров обусловлено тем, что они отражают два ключевых аспекта почвенных условий: влагообеспеченность растений и обеспеченность основными элементами минерального питания. Запас продуктивной влаги является одним из важных факторов, влияющих на формирование урожайности пшеницы, поскольку определяет доступность влаги для растений в разные фазы вегетации. Дефицит доступной влаги в период активного роста, колошения и налива зерна может приводить к снижению

продуктивности посевов. В научной литературе водный стресс рассматривается как одно из основных абиотических ограничений урожайности пшеницы, особенно в регионах с неустойчивым увлажнением. Для условий Северо-Казахстанской области учет влагообеспеченности имеет особое значение, поскольку урожайность зерновых культур в регионе в значительной степени зависит от количества осадков и запасов продуктивной влаги в почве. Поэтому в качестве входных параметров искусственной нейронной сети были использованы показатели запаса продуктивной влаги на нескольких глубинах. Такой подход позволяет учитывать не только поверхностное увлажнение, но и состояние более глубоких слоев почвы, которые могут иметь значение для корневой системы растений в засушливые периоды. Наряду с показателями запаса продуктивной влаги в модель включены содержание азота и фосфора. Азот и фосфор относятся к основным элементам минерального питания пшеницы и оказывают влияние на рост, развитие корневой системы, формирование вегетативной массы и продуктивность культуры. В исследованиях по минеральному питанию пшеницы подчеркивается, что фосфор способствует развитию корневой системы и начальному росту растений, а сбалансированное применение азота и фосфора рассматривается как один из управляемых факторов повышения урожайности.

Метеорологическая группа факторов  $WF$  включает количество осадков и среднюю температуру воздуха за май, июнь, июль и август. Указанные месяцы охватывают основной период вегетации яровой пшеницы в условиях Северо-Казахстанской области. Разделение метеорологических данных по месяцам позволяет учитывать не только общий уровень увлажнения и температурный режим, но и распределение погодных условий в течение вегетационного периода. Это важно, поскольку одинаковая сумма осадков за сезон может по-разному влиять на урожайность в зависимости от того, в какие фазы развития культуры она была получена.

Фитосанитарная группа факторов  $SF$  включает показатели сорности, распространения вредителей и болезней. В условиях Северного Казахстана посевам яровой пшеницы могут наносить ущерб хлебная полосатая блошка, серая зерновая совка, саранчовые и другие фитофаги. Существенное значение имеет и засоренность посевов: среди распространенных сорных растений встречаются овсюг, осот, вьюнок полевой, пырей ползучий, марь белая и другие виды. Сорная растительность конкурирует с культурными растениями за влагу, элементы питания, свет и пространство, что особенно важно в условиях недостаточного увлажнения. Распространение сорной растительности оценивалось по условной шкале от 0 до 4, где 0 соответствует отсутствию сорности, а 4 – очень высокому уровню ее распространения. Аналогичный подход применялся для оценки пораженности вредителями ( $X_{16}$ ): значение 0 соответствует отсутствию вредителей, а значение 4 – очень высокому уровню пораженности. Пораженность болезнями ( $X_{14}$ ) оценивалась по иной методике: в качестве показателя использовался суммарный индекс пораженности доминирующими заболеваниями пшеницы,

включая бурую листовую ржавчину, мучнистую росу и септориоз. Аддитивная природа данного индекса обуславливает его более широкий числовой диапазон по сравнению с показателями сорности и вредителей. Болезни пшеницы также могут снижать урожайность за счет поражения листьев, стеблей, корневой системы и колоса. В условиях региона к значимым заболеваниям относятся корневые гнили, бурая, желтая и стеблевая ржавчина, септориоз, гельминтоспориоз и другие пятнистости листьев [34]. При развитии заболеваний ухудшается физиологическое состояние растений, снижается площадь фотосинтезирующей поверхности, нарушается процесс налива зерна и уменьшается продуктивность посевов.

Данные по сорности, вредителям и болезням были включены в состав входных параметров искусственной нейронной сети, поскольку они отражают фитосанитарное состояние посевов и позволяют учитывать биологические ограничения формирования урожая.

Выходным параметром ИНС послужила урожайность яровой пшеницы (таблица 2.2).

Таблица 2.2 – Урожайность пшеницы яровой по Северо-Казахстанской области

Год	Районы Северо-Казахстанской области												
	Айыртаский	Акжарский	Жумабайев	Есилякский	Жамбылский	Кызылжарский	Мамлютский	Шалакына	Акжайынский	Тайыншинский	Тимирязевский	Уалихановский	Мусрепова
2007	14,6	14,2	12,8	14,7	14,8	15,7	14,3	16,9	15,6	15,0	16,0	13,1	16,1
2008	13,0	5,9	14,9	13,0	13,7	14,2	13,5	11,4	13,9	12,5	12,7	6,5	13,4
2009	14,1	13,0	17,1	15,7	13,6	16,8	15,7	13,1	17,6	13,9	14,4	13,1	12,7
2010	9,2	12,0	11,8	10,0	7,6	12,6	12,5	8,9	10,6	9,3	7,6	9,6	7,2
2011	23,0	15,8	21,9	23,4	19,4	22,8	21,0	19,1	21,7	19,1	25,5	14,6	22,5
2012	15,5	7,7	13,7	16,1	12,7	16,5	15,9	8,7	14,0	10,1	9,5	7,4	8,5
2013	12,6	9,1	15,2	11,5	11,0	13,7	11,2	9,1	12,8	12,5	9,4	12,3	14,5
2014	13,8	12,9	6,7	13,1	13,6	16,0	13,6	11,9	14,1	13,4	12,3	12,5	13,7
2015	14,8	15,0	18,3	17,7	17,3	18,7	15,9	13,2	18,2	12,9	14,4	15,1	14,5
2016	14,1	13,2	17,4	16,4	17,4	16,6	14,3	14,1	18,2	11,8	13,8	12,7	14,8
2017	16,8	12,6	19,9	18	18,8	17,4	17,4	15,9	19,8	14,6	17,3	11,6	16,2
2018	13,1	14,2	16,7	17	19,8	16,9	14,6	14,3	16,9	13,9	15,6	15,3	14,6
2019	12,7	12,4	13,5	17,3	15,4	18,2	17,7	14,1	16,3	14,2	11,8	14,5	13,2
2020	14,0	13,2	14,0	18,1	14,3	18,3	17,9	15,5	16,1	12,4	14,5	10,2	14,0

Данные, представленные в таблице 2.2, взяты с официального сайта Бюро национальной статистики Агентства по стратегическому планированию и реформам Республики Казахстан [35] и отражают фактические значения урожайности яровой пшеницы по 13 районам Северо-Казахстанской области за период 2007–2020 гг. Указанные значения использованы в исследовании в качестве выходного параметра искусственной нейронной сети, то есть целевой переменной  $y$ , прогнозируемой на основе совокупности входных агрометеорологических, почвенно-агрехимических и фитосанитарных факторов.

Анализ таблицы показывает, что урожайность пшеницы характеризуется выраженной пространственной и межгодовой изменчивостью. Наименьшее

значение урожайности за рассматриваемый период зафиксировано в Акжарском районе в 2008 г. и составило 5,9 ц/га, наибольшее значение – в Тимирязевском районе в 2011 г., где урожайность достигла 25,5 ц/га. Среднее значение урожайности по всей совокупности наблюдений составляет около 14,4 ц/га.

Наиболее низкие средние значения урожайности по области наблюдались в 2010, 2012 и 2013 гг., что указывает на неблагоприятные условия формирования урожая в данные годы. Напротив, 2011 г. характеризуется наиболее высоким уровнем урожайности практически во всех районах области. Такая изменчивость подтверждает необходимость учета не только районных различий, но и погодных, почвенных и фитосанитарных условий конкретного года.

## **2.2 Разведочный анализ данных (EDA) для построения нейросетевой модели прогнозирования урожайности**

EDA-анализ (Exploratory Data Analysis, разведочный анализ данных) выполнялся с применением языка программирования Python и специализированных библиотек – Pandas [36], Matplotlib [37] и Seaborn [38], обеспечивающих богатый инструментарий для исследовательской визуализации данных. Построение и обучение нейросетевой модели осуществлялись в среде MATLAB R2023b с применением Deep Learning Toolbox, располагающего специализированными алгоритмами для искусственных нейронных сетей – в том числе алгоритмом Левенберга–Марквардта и встроенными средствами кросс-валидации. Разделение инструментальных сред является стандартной практикой в прикладных исследованиях и обусловлено использованием оптимального инструмента для каждого этапа работы. Согласно Тьюки [39], основоположнику EDA, предварительный анализ данных позволяет выявить их структуру, аномальные наблюдения, характер распределений и направление взаимосвязей между переменными до применения каких-либо прогностических моделей. Целью EDA является углубленное понимание структуры исходных данных и обнаружение любых закономерностей, которые могут способствовать построению более точной и эффективной нейросетевой модели [40, 41].

До начала анализа проведена очистка и подготовка данных с целью обеспечения их корректности и достоверности. Данный шаг включал выявление пропущенных значений и устранение физически недопустимых наблюдений или статистически аномальных наблюдений.

Для первичного выявления возможных выбросов использовался метод межквартильного размаха IQR (англ. Interquartile Range). Данный метод относится к устойчивым статистическим подходам, поскольку основан на квартилях и менее чувствителен к экстремальным значениям по сравнению с методами, использующими среднее значение и стандартное отклонение.

Межквартильный размах определяется как разность между третьим и первым квартилями выборки:

$$IQR = Q_3 - Q_1 \quad (2.2)$$

где  $Q_1$  – первый квартиль, соответствующий 25-му процентилю выборки;  
 $Q_3$  – третий квартиль, соответствующий 75-му процентилю выборки.

На основе значения IQR рассчитываются нижняя и верхняя границы допустимого диапазона значений:

$$X_{min}^* = Q_1 - 1,5 \cdot IQR, \quad (2.3)$$

$$X_{max}^* = Q_3 + 1,5 \cdot IQR, \quad (2.4)$$

Значения признака, расположенные ниже  $X_{min}^*$  или выше  $X_{max}^*$ , рассматриваются как потенциальные выбросы и требуют дополнительной содержательной проверки. В рамках настоящего исследования такие значения не исключались автоматически, поскольку в аграрных данных экстремальные наблюдения могут отражать реальные особенности погодных, почвенных или фитосанитарных условий конкретного года. Поэтому метод IQR использовался как инструмент первичной диагностики качества данных перед последующим нейросетевым моделированием [42].

Для целевой переменной  $Y$ , представляющей урожайность пшеницы, расчет был выполнен по выборке  $n=182$ . По результатам расчета первый квартиль составил  $Q_1=12,70$  ц/га, третий квартиль –  $Q_3=16,27$  ц/га, межквартильный размах –  $IQR=3,57$  ц/га. Соответственно, нижняя граница допустимого диапазона составила:

$$12,70 - 1,5 \cdot 3,57 = 7,34 \text{ ц/га,}$$

а верхняя граница:

$$16,27 + 1,5 \cdot 3,57 = 21,64 \text{ ц/га.}$$

За пределы верхней границы вышли 7 наблюдений, что составляет 3,8% выборки. Их значения находятся в диапазоне от 21,7 до 25,5 ц/га. Все указанные наблюдения относятся к районам Северо-Казахстанской области в 2011 году: Аккайынский район – 21,7 ц/га, район имени М. Жумабаева – 21,9 ц/га, район имени Г. Мусрепова – 22,5 ц/га, Кызылжарский район – 22,8 ц/га, Айыртауский район – 23,0 ц/га, Есильский район – 23,4 ц/га, Тимирязевский район – 25,5 ц/га.

За пределы нижней границы вышли 4 наблюдения, что составляет 2,2% выборки. Их значения находятся в диапазоне от 5,9 до 7,2 ц/га и относятся к отдельным районам в неблагоприятные по урожайности 2008 и 2010 годы.

Значения урожайности, выявленные методом IQR как статистические выбросы, не были исключены из исходного набора данных. Такое решение

обусловлено тем, что указанные наблюдения не являются техническими ошибками, а отражают реальные экстремальные агрономические условия. Высокие значения урожайности 2011 года и низкие значения урожайности 2008 и 2010 годов характеризуют фактическую межгодовую изменчивость продуктивности пшеницы в районах Северо-Казахстанской области. Удаление таких наблюдений могло бы привести к снижению способности модели учитывать как благоприятные, так и неблагоприятные условия формирования урожая.

При этом в ходе проверки методом IQR и логической проверки диапазонов были выявлены отдельные технически некорректные значения в метеорологических переменных. В частности, в переменной  $X_{12}$ , характеризующей температуру воздуха в июле, было зафиксировано значение 112,4 °С, а в переменной  $X_{13}$ , характеризующей температуру воздуха в августе, – значение 61,1 °С. Указанные значения являются физически невозможными для условий исследуемого региона и были отнесены к техническим ошибкам ввода данных. После сопоставления с данными смежных метеорологических наблюдений они были скорректированы соответственно до 11,2 °С и 21,1 °С.

При подготовке данных были разграничены два типа аномальных значений. К первой группе отнесены реальные экстремальные наблюдения целевой переменной, отражающие особенности конкретных агроклиматических лет; они были сохранены в исходном наборе данных. Ко второй группе отнесены физически невозможные значения метеорологических признаков, которые были идентифицированы как технические ошибки и скорректированы до последующего моделирования [7,8]. Такой подход позволил сохранить информативность выборки и одновременно исключить влияние ошибочных записей на процесс обучения искусственной нейронной сети.

После очистки и верификации исходного набора данных был выполнен расчет описательных статистик по каждой переменной. Данный этап необходим для первичной оценки диапазона изменения признаков, степени их вариабельности, а также характера линейной связи между входными параметрами и целевой переменной – урожайностью пшеницы. Для этого были определены минимальное и максимальное значения, среднее значение, стандартное отклонение, коэффициент вариации и коэффициент линейной корреляции Пирсона.

Среднее значение признака рассчитывалось по формуле:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (2.5)$$

где  $x_i$  – значение признака в  $i$ -м наблюдении;

$n$  – количество наблюдений.

Стандартное отклонение  $\sigma$ , характеризующее разброс значений относительного среднего, определялось по формуле:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.6)$$

Для оценки относительной изменчивости признаков использовался коэффициент вариации:

$$CV = \frac{\sigma}{\bar{x}} \cdot 100\% \quad (2.7)$$

Коэффициент линейной корреляции Пирсона между входным признаком  $X$  и целевой переменной  $Y$  рассчитывался по формуле:

$$r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (2.8)$$

где  $y_i$  – значение урожайности в  $i$ -м наблюдении;  
 $\bar{y}$  – среднее значение урожайности.

Коэффициент корреляции Пирсона использовался для предварительной оценки направления и силы линейной связи между отдельными входными признаками и урожайностью. При этом результаты корреляционного анализа рассматривались только как элемент разведочного анализа данных и не интерпретировались как доказательство причинно-следственного влияния факторов на урожайность.

Сводные результаты описательной статистики по 17 переменным, включая 16 входных признаков и целевую переменную, представлены в таблице 2.3.

Таблица 2.3– Описательная статистика переменных нейросетевой модели

Переменная	Мин.	Макс.	Среднее	$\sigma$	CV, %	$r(X,Y)$
Y (урожайность, ц/га)	5,9	25,5	14,52	3,43	23,6	—
X1 (запас прод. влаги 0–10 см, мм)	2,4	27,6	11,09	4,22	38,1	+0,057
X2 (запас прод. влаги 0–50 см, мм)	22,6	109,6	56,07	16,4	29,3	+0,162
X3 (запас прод. влаги 0–100 см, мм)	47,3	191,3	111,5	27,5	24,7	+0,122
X4 (азот нитратный, мг/кг)	31,9	92,7	49,56	5,95	12,0	+0,133
X5 (фосфор подвижный, мг/кг)	6,1	53,3	18,70	5,00	26,7	+0,023
X6 (осадки май, мм)	4,0	100,9	38,58	19,9	51,7	+0,317

X <sub>7</sub> (осадки июнь, мм)	3,3	156,4	36,95	24,7	66,9	+0,339
X <sub>8</sub> (осадки июль, мм)	5,1	285,0	72,94	38,1	52,3	+0,060
X <sub>9</sub> (осадки август, мм)	4,9	145,2	43,49	25,3	58,2	-0,003
X <sub>10</sub> (Т май, °С)	6,9	16,5	13,08	2,06	15,8	-0,143
X <sub>11</sub> (Т июнь, °С)	12,5	23,2	18,12	1,90	10,5	-0,205
X <sub>12</sub> (Т июль, °С)	11,2	24,5	18,88	1,89	10,0	-0,250
X <sub>13</sub> (Т август, °С)	14,0	22,3	17,67	1,69	9,6	-0,531
X <sub>14</sub> (индекс болезней, у.е.)	100,9	128,0	122,6	1,67	1,4	+0,110
X <sub>15</sub> (сорность, у.е.)	1,5	3,0	1,92	0,16	8,4	-0,254
X <sub>16</sub> (вредители, у.е.)	0,0	4,0	1,54	0,51	33,0	-0,084

В таблице 2.3 приведены результаты описательной статистики переменных, использованных при построении нейросетевой модели прогнозирования урожайности пшеницы. В составе таблицы представлены целевая переменная  $Y$ , отражающая урожайность пшеницы, а также 16 входных признаков  $X_1$ – $X_{16}$ , характеризующих агрохимические, метеорологические и фитосанитарные условия формирования урожая.

В графах «Мин.» и «Макс.» указаны минимальные и максимальные значения переменных в исследуемой выборке. Графа «Среднее» отражает среднее арифметическое значение показателя,  $\sigma$  – стандартное отклонение, характеризующее степень разброса значений относительно среднего уровня. Коэффициент вариации  $CV$  представлен в процентах и используется для оценки относительной изменчивости признаков. Последний столбец  $r(X, Y)$  содержит коэффициент линейной корреляции Пирсона между соответствующим входным признаком  $X_i$  и целевой переменной  $Y$ , то есть урожайностью пшеницы.

Анализ целевой переменной показывает, что средняя урожайность пшеницы за исследуемый период составляет 14,52 ц/га. Минимальное значение урожайности равно 5,9 ц/га, максимальное – 25,5 ц/га. Стандартное отклонение составляет 3,43 ц/га, а коэффициент вариации – 23,6%. Полученные значения свидетельствуют о наличии умеренной изменчивости урожайности, обусловленной пространственными различиями между районами Северо-Казахстанской области и межгодовой вариацией агроклиматических условий.

Показатели почвенной влаги  $X_1$ – $X_3$  характеризуются различным уровнем изменчивости. Наибольший коэффициент вариации в данной группе наблюдается у

показателя  $X_1$ , соответствующего слою 0–10 см ( $CV = 38,1\%$ ). Для показателей  $X_2$  и  $X_3$ , характеризующих более глубокие слои почвы, коэффициент вариации составляет 29,3% и 24,7% соответственно. Это указывает на то, что верхний слой почвы отличается более высокой динамичностью, тогда как значения влаги в более глубоких слоях изменяются относительно менее резко.

Агрохимические показатели представлены содержанием нитратного азота  $X_4$  и подвижного фосфора  $X_5$ . Коэффициент вариации нитратного азота составляет 12,0%, что указывает на сравнительно низкую изменчивость данного признака. Для подвижного фосфора коэффициент вариации равен 26,7%, что свидетельствует о более выраженной неоднородности данного показателя в исследуемой выборке.

Метеорологические параметры демонстрируют существенные различия по уровню изменчивости. Температурные показатели  $X_{10}$ – $X_{13}$  имеют относительно низкие значения коэффициента вариации – от 9,6% до 15,8%. В то же время показатели осадков  $X_6$ – $X_9$  характеризуются значительно более высокой вариабельностью: от 51,7% для осадков мая до 66,9% для осадков июня. Это подтверждает высокую межгодовую неустойчивость атмосферного увлажнения в период вегетации пшеницы.

Фитосанитарные показатели  $X_{14}$ – $X_{16}$  также имеют неодинаковую степень изменчивости. Индекс болезней  $X_{14}$  характеризуется низким коэффициентом вариации ( $CV = 1,4\%$ ), показатель сорности  $X_{15}$  –  $CV = 8,4\%$ . При этом показатель вредителей  $X_{16}$  изменяется в диапазоне от 0,0 до 4,0 у.е. и имеет коэффициент вариации 33,0%, что свидетельствует о более выраженной неоднородности проявления данного фактора по годам и районам.

Корреляционный анализ показывает, что наиболее выраженная линейная связь с урожайностью выявлена для температуры воздуха в августе  $X_{13}$ , где коэффициент корреляции составляет  $r(X, Y) = -0,531$ . Отрицательное значение коэффициента указывает на обратную связь между данным признаком и урожайностью в исследуемой выборке. Также отрицательная корреляционная связь наблюдается для сорности  $X_{15}$  ( $r = -0,254$ ), температуры июля  $X_{12}$  ( $r = -0,250$ ) и температуры июня  $X_{11}$  ( $r = -0,205$ ).

Среди признаков, имеющих положительную связь с урожайностью, наиболее заметные значения коэффициента корреляции получены для осадков июня  $X_7$  ( $r = +0,339$ ) и осадков мая  $X_6$  ( $r = +0,317$ ). Это указывает на то, что в рассматриваемой выборке увеличение количества осадков в мае и июне связано с более высокими значениями урожайности. Остальные признаки характеризуются слабой линейной связью с целевой переменной.

Следует отметить, что коэффициенты корреляции, представленные в таблице 2.3, используются исключительно как инструмент разведочного анализа данных. Они позволяют предварительно оценить направление и силу линейной связи между признаками и урожайностью, однако не являются доказательством причинно-следственного влияния отдельных факторов на формирование урожая. Данные различия в масштабах переменных обуславливают необходимость нормализации перед подачей в нейронную сеть.

Для визуального анализа распределения целевой переменной была построена гистограмма урожайности, а также диаграмма размаха (boxplot) по выделенным типам агроклиматических лет. Результаты представлены на рисунке 2.2.

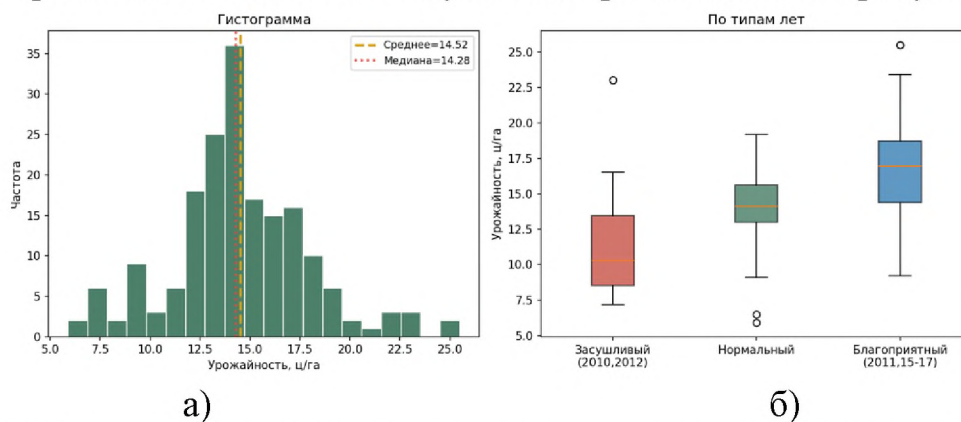


Рисунок 2.2 – Распределение урожайности яровой пшеницы по СКО (2007–2020 гг.):

- а) гистограмма с линиями среднего и медианы;
- б) ящик с усами по типам агроклиматических лет

На рисунке 2.2а представлена гистограмма распределения урожайности яровой пшеницы по всей совокупности наблюдений. Вертикальными пунктирными линиями отмечены среднее значение урожайности и медиана. Среднее значение составляет 14,52 ц/га, медиана – 14,28 ц/га. Близость среднего значения и медианы указывает на отсутствие резко выраженной асимметрии распределения целевой переменной. Основная часть наблюдений сосредоточена в интервале примерно от 12 до 17 ц/га, что соответствует наиболее типичным значениям урожайности в исследуемой выборке.

Вместе с тем гистограмма показывает наличие как относительно низких, так и относительно высоких значений урожайности. Низкие значения отражают неблагоприятные условия формирования урожая в отдельные годы и районы, тогда как высокие значения связаны с благоприятным сочетанием погодных, почвенных и фитосанитарных условий. Таким образом, распределение урожайности

характеризуется естественной вариативностью, обусловленной пространственными различиями между районами и межгодовой изменчивостью агроклиматических условий.

На рисунке 2.2б представлена диаграмма размаха или «ящик с усами» (boxplot), позволяющая сравнить распределение урожайности по типам агроклиматических лет. В диаграмме размаха центральная линия внутри прямоугольника соответствует медиане, нижняя и верхняя границы прямоугольника – первому и третьему квартилям, а высота прямоугольника отражает межквартильный размах. «Усы» диаграммы показывают диапазон значений, не относящихся к потенциальным выбросам, а отдельные точки за пределами «усов» обозначают наблюдения, отличающиеся от основной части распределения.

Анализ диаграммы размаха показывает, что медианное значение урожайности возрастает при переходе от засушливых лет к нормальным и далее к благоприятным. Для засушливых лет характерны более низкие значения урожайности и наличие отдельных наблюдений, выходящих за пределы основной части распределения. Нормальные годы занимают промежуточное положение как по уровню медианы, так и по разбросу значений. Благоприятные годы характеризуются более высоким уровнем урожайности и смещением распределения в сторону больших значений.

Отдельные точки, отображенные на диаграмме размаха, не следует автоматически рассматривать как ошибочные наблюдения. В аграрных данных такие значения могут отражать реальные экстремальные условия конкретного года или района. Поэтому они подлежат содержательной интерпретации и верификации, а не механическому исключению из выборки.

Для оценки временной изменчивости целевой переменной была построена динамика средней урожайности яровой пшеницы по районам Северо-Казахстанской области за период 2007–2020 гг. Данный график позволяет проследить изменение урожайности по годам, а также оценить разброс значений между районами внутри каждого года.

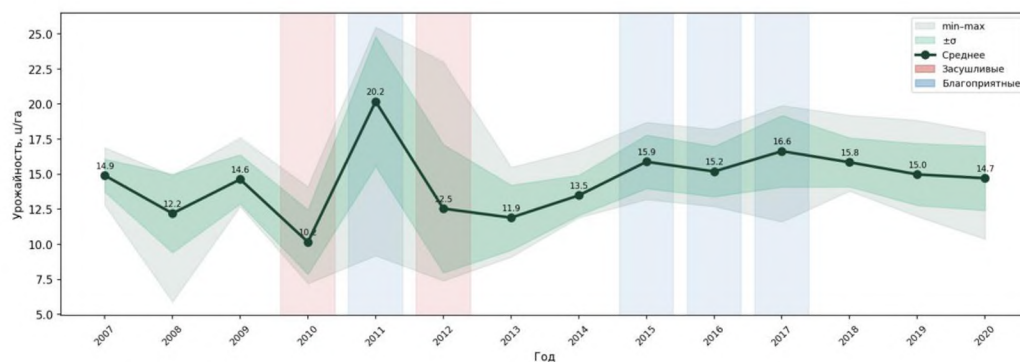


Рисунок 2.3 – Динамика средней урожайности яровой пшеницы по СКО за 2007–2020 гг.

На рисунке 2.3 представлена межгодовая динамика средней урожайности яровой пшеницы по 13 районам Северо-Казахстанской области. Сплошная линия с маркерами отражает среднее значение урожайности, рассчитанное по совокупности районов за каждый год наблюдения. Зеленая область соответствует интервалу  $\pm\sigma$ , характеризующему разброс значений относительно среднего уровня, а светлая область показывает диапазон между минимальным и максимальным значениями урожайности среди районов области.

Цветовая заливка использована для визуального выделения лет, отличающихся от основной части временного ряда. Красным цветом обозначены годы с пониженной средней урожайностью, голубым – годы с повышенной средней урожайностью. В качестве ориентира использовались долгосрочное среднее значение урожайности по выборке, равное 14,52 ц/га, и стандартное отклонение  $\sigma=3,43$  ц/га. При этом 2010 г. характеризуется наиболее выраженным снижением урожайности, превышающим одно стандартное отклонение относительно долгосрочного среднего. К периоду пониженной урожайности также отнесен 2012 г., однако его отклонение от среднего уровня не достигает порога  $-1\sigma$ , поэтому данный год рассматривается как год с умеренно сниженной урожайностью. Наиболее выраженное повышение урожайности зафиксировано в 2011 г., когда среднее значение достигло 20,2 ц/га. Годы 2015–2020 характеризуются относительно повышенным и более устойчивым уровнем урожайности по сравнению с предыдущим периодом.

Анализ представленной динамики свидетельствует о выраженной межгодовой изменчивости урожайности пшеницы. Минимальное среднее значение за исследуемый период наблюдалось в 2010 г. и составило около 10,1 ц/га. Относительно низкий уровень урожайности также отмечен в 2008, 2012 и 2013 гг. Наибольшее среднее значение урожайности зафиксировано в 2011 г. Начиная с 2015 г., средняя урожайность находилась преимущественно в диапазоне от 14,7 до 16,6 ц/га, что указывает на более устойчивый характер значений в заключительной части рассматриваемого периода.

Ширина области min–max в отдельные годы указывает на наличие существенных различий между районами Северо-Казахстанской области. Это свидетельствует о том, что формирование урожайности определяется не только общими условиями конкретного года, но и пространственными особенностями районов, включая уровень влагообеспеченности почвы, агрохимическое состояние и фитосанитарную обстановку. Следовательно, при построении нейросетевой модели необходимо учитывать, как временную, так и пространственную неоднородность исходных данных.

Для предварительной оценки линейных взаимосвязей между переменными была построена корреляционная матрица. В качестве меры связи использовался коэффициент линейной корреляции Пирсона  $r_{XY}$ , который является классической статистической характеристикой направления и силы линейной зависимости между двумя переменными [43]. Корреляционная матрица представляет собой квадратную матрицу, элементы которой содержат парные коэффициенты корреляции между анализируемыми переменными; такой подход широко применяется в прикладном многомерном анализе данных [44]. Визуальное представление корреляционной матрицы выполнено в виде тепловой карты, где цвет ячейки отражает направление и степень выраженности линейной связи между входными признаками, а также между отдельными признаками и целевой переменной – урожайностью пшеницы.

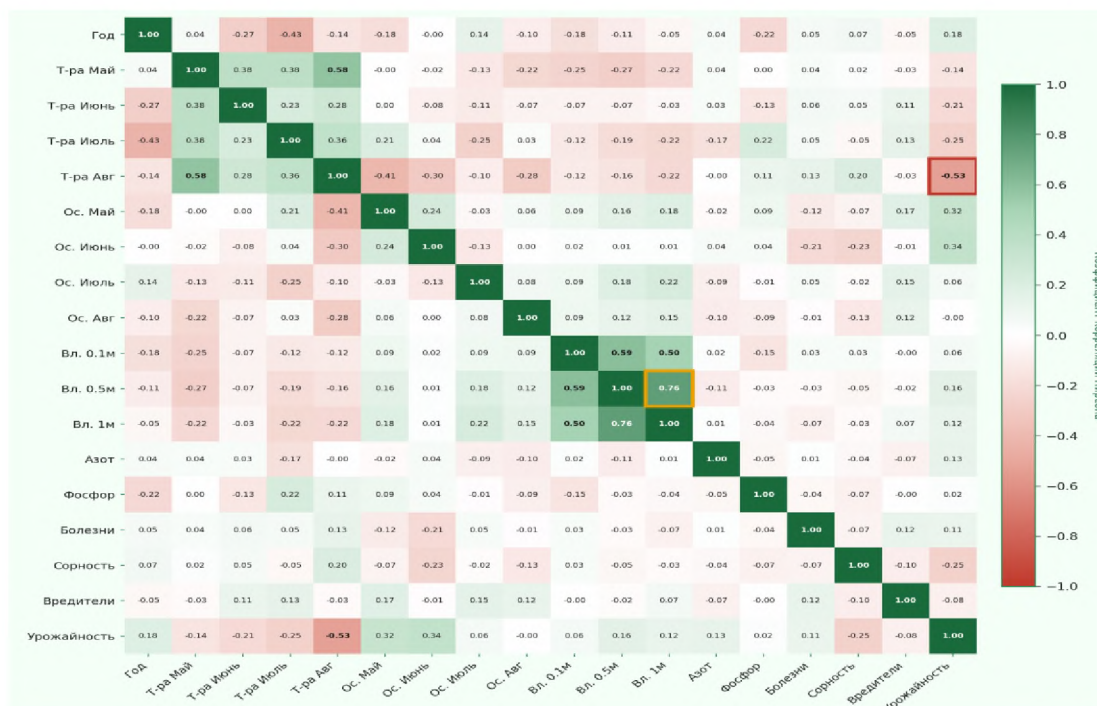


Рисунок 2.4 – Тепловая карта корреляций между входными признаками и урожайностью пшеницы

На рисунке 2.4 представлена тепловая карта корреляций между временным показателем, входными признаками нейросетевой модели и целевой переменной.

Зеленые оттенки соответствуют положительным значениям коэффициента корреляции, красные – отрицательным. Чем выше интенсивность цвета, тем сильнее выражена линейная связь между соответствующими переменными.

Анализ корреляционной матрицы показывает, что наиболее выраженная отрицательная связь с урожайностью наблюдается у температуры воздуха в августе. Коэффициент корреляции между температурой августа и урожайностью составляет  $r=-0,53$ , что указывает на умеренную обратную линейную связь. Это означает, что в рассматриваемой выборке повышение температуры воздуха в августе связано со снижением урожайности пшеницы. Также отрицательная связь с урожайностью отмечается для температуры июля ( $r=-0,25$ ), температуры июня ( $r=-0,21$ ) и показателя сорности ( $r=-0,25$ ).

Среди положительных корреляционных связей с урожайностью наиболее заметными являются связи с осадками июня ( $r=+0,34$ ) и осадками мая ( $r=+0,32$ ). Это свидетельствует о том, что увеличение количества осадков в мае и июне в исследуемой выборке связано с более высокими значениями урожайности. Однако связи имеют слабый и умеренный характер, что указывает на невозможность объяснения урожайности только одним фактором.

Корреляционная матрица также показывает наличие взаимосвязей между отдельными входными признаками. Наиболее выраженная положительная связь выявлена между влажностью почвы на глубинах 0,5 м и 1 м ( $r=+0,76$ ). Это закономерно, поскольку данные показатели характеризуют состояние влагообеспеченности почвенного профиля на смежных глубинах. Кроме того, положительная связь наблюдается между влажностью почвы на глубинах 0,1 м и 0,5 м ( $r=+0,59$ ), а также между влажностью почвы на глубинах 0,1 м и 1 м ( $r=+0,50$ ).

Отдельные метеорологические признаки также демонстрируют взаимосвязи между собой. Например, температура воздуха в мае имеет положительную связь с температурой августа ( $r=+0,58$ ), что отражает общую согласованность температурного режима в пределах вегетационного периода. В то же время между температурой августа и осадками мая наблюдается отрицательная связь ( $r=-0,41$ ), что может указывать на различие погодных условий в отдельные годы.

Следует отметить, что показатель «Год» включен в корреляционную матрицу только для оценки временной направленности данных и не рассматривается как агробиологический фактор формирования урожайности. Его связь с урожайностью является слабой ( $r=+0,18$ ), что подтверждает отсутствие простой линейной тенденции роста или снижения урожайности за весь период наблюдений.

Полученные результаты показывают, что отдельные признаки имеют различное направление и силу связи с урожайностью. При этом большинство корреляций с целевой переменной не являются высокими по абсолютному значению. Это указывает на сложный, многомерный и вероятно нелинейный характер зависимости урожайности пшеницы от совокупности почвенных, метеорологических, агрохимических и фитосанитарных факторов.

После построения корреляционной матрицы было выполнено ранжирование входных переменных по абсолютному значению коэффициента корреляции Пирсона с целевой переменной. Такой подход позволяет выделить признаки, имеющие наиболее выраженную линейную связь с урожайностью пшеницы, независимо от направления этой связи. При этом знак коэффициента корреляции сохраняется: положительные значения отражают прямую связь с урожайностью, отрицательные – обратную.

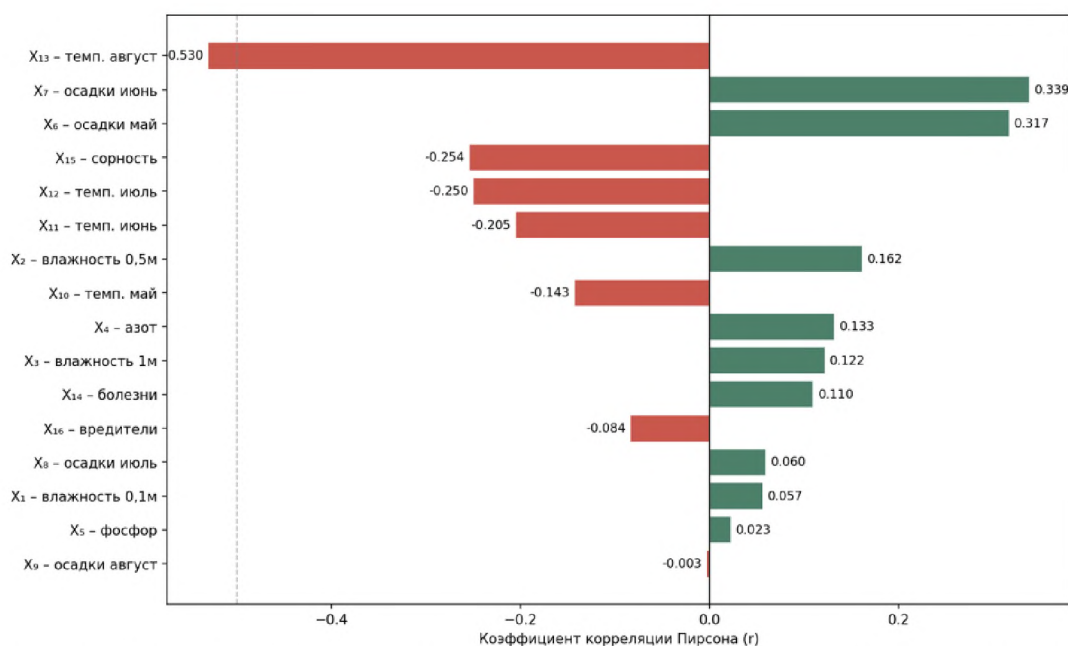


Рисунок 2.5 – Ранжирование входных переменных по абсолютному значению коэффициента корреляции Пирсона с урожайностью пшеницы

Анализ рисунка 2.5 показывает, что наиболее выраженная по модулю корреляционная связь с урожайностью наблюдается у температуры воздуха в августе  $X_{13}$ , для которой коэффициент корреляции составляет  $r=-0,530$ . Отрицательный знак указывает на обратную связь: в исследуемой выборке повышение августовской температуры связано со снижением урожайности пшеницы.

Наиболее заметные положительные связи выявлены для осадков июня  $X_7$  ( $r=+0,339$ ) и осадков мая  $X_6$  ( $r=+0,317$ ). Это свидетельствует о том, что увеличение количества осадков в первой половине вегетационного периода связано с более высокими значениями урожайности. Данный результат согласуется с агрономической логикой формирования урожая яровой пшеницы, поскольку влагообеспеченность в период активного роста растений является одним из важных условий продуктивности.

Среди факторов, имеющих отрицательную связь с урожайностью, кроме температуры августа, выделяются сорность  $X_{15}$  ( $r=-0,254$ ), температура июля  $X_{12}$  ( $r=-0,250$ ) и температура июня  $X_{11}$  ( $r=-0,205$ ). Отрицательная корреляция данных

признаков указывает на то, что их увеличение в рассматриваемой выборке связано со снижением урожайности.

Остальные переменные характеризуются более слабыми линейными связями с целевой переменной. В частности, показатели влажности почвы, содержания азота, фосфора, болезней и вредителей имеют коэффициенты корреляции меньшей величины. Это не означает, что данные признаки не оказывают влияния на урожайность, поскольку корреляционный анализ отражает только линейные зависимости и не учитывает возможные нелинейные взаимодействия между факторами.

Рисунок 2.5 позволяет выделить наиболее информативные с точки зрения линейной связи признаки: температуру воздуха в августе, осадки июня и мая, сорность, а также температуру июля и июня. Вместе с тем умеренные и слабые значения большинства коэффициентов корреляции подтверждают, что урожайность пшеницы формируется под воздействием совокупности факторов, а не одного изолированного показателя. Это дополнительно обосновывает применение нейросетевой модели, способной учитывать сложные многомерные зависимости между входными параметрами и целевой переменной.

После построения корреляционной матрицы и ранжирования входных признаков по величине связи с урожайностью были построены диаграммы рассеяния для наиболее значимых предикторов. Для визуального анализа выбраны четыре переменные, имеющие наиболее выраженную корреляционную связь с целевой переменной: температура воздуха в августе  $X_{13}$ , осадки июня  $X_7$ , осадки мая  $X_6$  и сорность  $X_{15}$ . Значения коэффициента корреляции Пирсона, указанные на графиках, используются для оценки направления и силы линейной связи между отдельным признаком и урожайностью; данный коэффициент является классической мерой линейной зависимости между двумя переменными [45].

Для дополнительной визуальной интерпретации на диаграммах рассеяния построены линии линейного тренда. В общем виде зависимость урожайности от отдельного входного признака может быть представлена следующим образом:

$$\hat{Y}_j = a + bx_{ij}, \quad (2.9)$$

где  $\hat{Y}_j$  – расчетное значение урожайности для  $j$ -го наблюдения;

$x_{ij}$  – значение  $i$ -го входного признака в  $j$ -м наблюдении;

$a$  – свободный член линейной модели;

$b$  – коэффициент наклона прямой.

Параметры  $a$  и  $b$  определяются методом наименьших квадратов, при котором минимизируется сумма квадратов отклонений фактических значений от расчетных. Метод наименьших квадратов является стандартным способом оценки параметров

линейной регрессии и основан на минимизации суммы квадратов остатков. В данном случае целевая функция имеет вид:

$$S_{(a,b)} = \sum_{j=1}^n (y_j - a - bx_{ij})^2 \rightarrow \min, \quad (2.10)$$

где  $y_j$  – фактическое значение урожайности в  $j$ -м наблюдении;  
 $n$  – количество наблюдений.

Коэффициент наклона прямой рассчитывается по формуле:

$$b = \frac{\sum_{j=1}^n (x_{ij} - \bar{x}_i)(y_j - \bar{y})}{\sum_{j=1}^n (x_{ij} - \bar{x}_i)^2}, \quad (2.11)$$

а свободный член определяется как:

$$a = \bar{y} - b\bar{x}_i \quad (2.12)$$

Линии тренда на рисунке 2.6 используются только как средство визуального отображения направления линейной связи между отдельным признаком и урожайностью. Они не рассматриваются как самостоятельные прогнозные модели, поскольку урожайность пшеницы формируется под влиянием совокупности факторов.

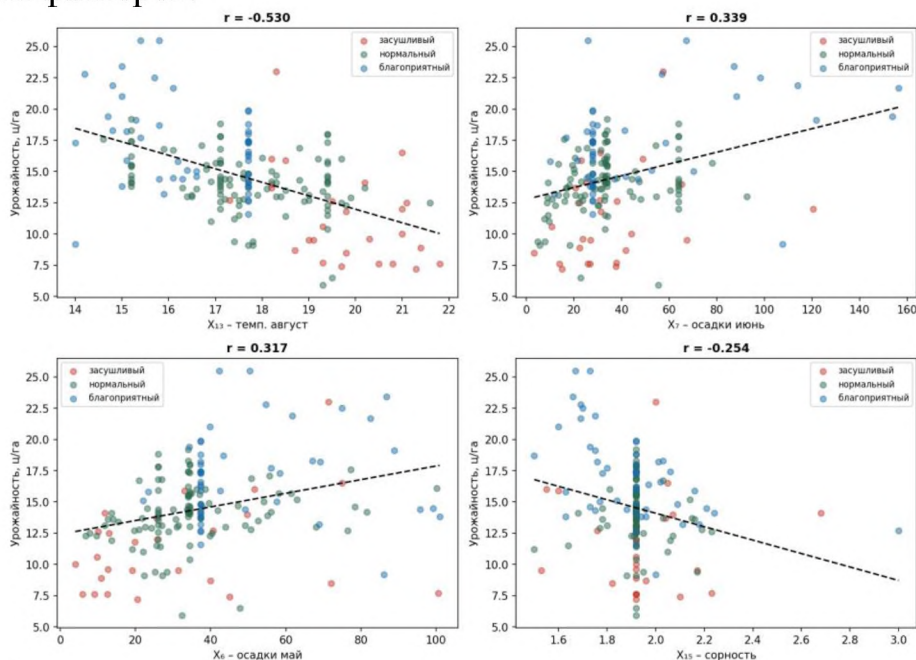


Рисунок 2.6 – Взаимосвязь ключевых входных признаков с урожайностью пшеницы по данным 2007–2020 гг.

Анализ рисунка 2.6 визуально согласуется с результатами корреляционного анализа. Среди рассмотренных предикторов наиболее выраженная обратная линейная связь с урожайностью наблюдается для среднемесячной температуры воздуха в августе ( $X_{13}$ ), для которой коэффициент корреляции составляет  $r=-0,531$ . Наклон линии тренда показывает, что в исследуемой выборке повышение августовской температуры связано со снижением урожайности пшеницы.

Положительная связь выявлена между урожайностью и количеством осадков в июне  $X_7$  ( $r=+0,339$ ), а также в мае  $X_6$  ( $r=+0,317$ ). Это указывает на то, что увеличение количества осадков в первой половине вегетационного периода в рассматриваемой совокупности наблюдений связано с более высокими значениями урожайности. Такой результат согласуется с агрономической логикой, поскольку влагообеспеченность в начальные и активные фазы развития яровой пшеницы является одним из важных условий формирования продуктивности. Для показателя сорности  $X_{15}$  наблюдается отрицательная связь с урожайностью ( $r=-0,254$ ). Это означает, что увеличение уровня сорности в исследуемой выборке связано со снижением урожайности пшеницы. Такая зависимость объясняется конкуренцией сорной растительности с культурными растениями за влагу, элементы питания, свет и пространство.

Цветовое разделение точек по типам лет позволяет дополнительно оценить распределение наблюдений в зависимости от общего уровня урожайности. Наблюдения, относящиеся к годам с повышенной урожайностью, чаще расположены в верхней части графиков, тогда как годы с пониженной урожайностью сосредоточены преимущественно в области более низких значений урожайности. При этом значительный разброс точек показывает, что связь между отдельными признаками и урожайностью не является однозначной и полностью линейной.

Следовательно, диаграммы рассеяния показывают, что отдельные входные признаки действительно связаны с урожайностью, однако ни один из них не объясняет ее изменение полностью. Это подтверждает многомерный характер задачи прогнозирования и обосновывает применение нейросетевой модели, способной учитывать совместное и потенциально нелинейное влияние агрометеорологических, почвенных, агрохимических и фитосанитарных факторов.

### **2.3 Предварительная обработка данных для нейросетевого моделирования**

После проведения разведочного анализа исходные данные были подготовлены к последующему нейросетевому моделированию. Предварительная обработка данных включала формирование входной матрицы признаков, выделение целевой переменной, проверку корректности числового формата, нормализацию переменных и подготовку массива данных к обучению искусственной нейронной сети в среде MATLAB.

Исходный набор данных был представлен в виде таблицы, где каждая строка соответствует одному наблюдению типа «район–год», а столбцы содержат значения входных признаков и целевой переменной. Всего в исследовании использовано 182 наблюдения, сформированные по 13 районам Северо-Казахстанской области за период 2007–2020 гг. Для каждого наблюдения были заданы 16 входных признаков  $X_1$ – $X_{16}$  и одно выходное значение  $Y$ , соответствующее урожайности яровой пшеницы.

Входной вектор модели имеет вид:

$$X = \{X_1, X_2, \dots, X_{16}\}, \quad (2.13)$$

где  $X_1$  –  $X_{16}$  – входные признаки, характеризующие агрохимические, метеорологические и фитосанитарные условия формирования урожая.

Целевая переменная задается как:

$$Y = \{y_1, y_2, \dots, y_n\}, \quad (2.14)$$

где  $y_1$  – фактическое значение урожайности пшеницы в  $n$ -м наблюдении;  
 $n = 182$  – общее количество наблюдений.

Для дальнейшего обучения нейронной сети данные были представлены в матричном виде. Матрица входных признаков содержит 182 строки и 16 столбцов:

$$X \in \mathbb{R}^{182 \times 16}, \quad (2.15)$$

а вектор целевой переменной имеет размерность:

$$Y \in \mathbb{R}^{182 \times 1}, \quad (2.16)$$

При передаче данных в MATLAB матрица входов была транспонирована в формат, используемый функциями нейросетевого моделирования:

$$\text{inputs} \in \mathbb{R}^{182 \times 16}, \quad \text{targets} \in \mathbb{R}^{1 \times 182}, \quad (2.17)$$

Такой формат означает, что строки матрицы `inputs` соответствуют входным признакам, а столбцы – отдельным наблюдениям. Вектор `targets` содержит фактические значения урожайности, соответствующие каждому наблюдению.

Перед нормализацией была выполнена проверка корректности данных. Проверялись отсутствие пропущенных значений, числовой формат всех переменных, отсутствие значений NaN и Inf, а также соответствие диапазонов

значений физическому смыслу признаков. В дальнейшем моделирование передавался верифицированный исходный набор данных после исправления технических ошибок, выявленных на этапе EDA-анализа. При этом реальные экстремальные значения урожайности, отражающие особенности отдельных лет и районов, были сохранены, поскольку они являются информативными для обучения модели в условиях межгодовой изменчивости урожайности.

Поскольку входные признаки имеют различные единицы измерения и числовые диапазоны, перед обучением была выполнена нормализация данных. В состав входного вектора входят показатели запаса продуктивной влаги, содержания азота и фосфора, количества осадков, температуры воздуха и фитосанитарные индексы. Непосредственное использование признаков, различающихся по единицам измерения и числовым диапазонам, может привести к тому, что переменные с большими абсолютными значениями будут оказывать непропорционально сильное влияние на процесс настройки весовых коэффициентов. Поэтому нормализация является обязательным этапом подготовки данных к обучению нейронной сети.

В настоящем исследовании применялась линейная нормализация методом *mapminmax*, реализованным в MATLAB Deep Learning Toolbox. Данный метод приводит значения переменных к диапазону  $[-1;+1]$ , что согласуется с областью значений функции активации *tansig*, использованной в скрытом слое нейронной сети. В общем виде нормализация выполнялась по формуле:

$$x_{norm} = 2 \cdot \frac{x - x_{min}}{x_{max} - x_{min}} - 1, \quad (2.18)$$

где  $x$  – исходное значение признака;

$x_{min}$  – минимальное значения данного признака;

$x_{max}$  – максимальное значения данного признака;

$x_{norm}$  – нормализованное значение признака.

Для наглядного представления результата масштабирования в таблице 2.4 приведен фрагмент исходных и нормализованных значений отдельных переменных. В качестве примера показаны признаки, имеющие разные единицы измерения и числовые диапазоны: запас продуктивной влаги  $X_1$ , количество осадков за май  $X_6$ , температура воздуха за август  $X_{13}$ , а также целевая переменная  $Y$ .

Таблица 2.4– Фрагмент данных до и после нормализации методом *mapminmax*

Район, год	$X_1$ , мм	$X_1$ норм.	$X_6$ , мм	$X_6$ норм.	$X_{13}$ , °C	$X_{13}$ норм.	$Y$ , ц/га	$Y$ норм.
Акжарский, 2008	8,9	-0,48	32,4	-0,41	19,3	+0,28	5,9	-1,00

Акжарский, 2010	27,6	+1,00	25,8	-0,55	21,0	+0,69	12,0	-0,38
Айыртауский, 2007	27,0	+0,95	78,4	+0,54	17,1	-0,25	14,6	-0,11
Аккайынский, 2017	10,9	-0,33	37,3	-0,31	17,7	-0,11	19,8	+0,42
Тимирязевский, 2011	10,4	-0,37	50,4	-0,04	15,8	-0,57	25,5	+1,00

Как видно из таблицы 2.4, исходные значения признаков имеют различные масштабы измерения, однако после применения нормализации методом  $\text{mapminmax}$  они приводятся к единому диапазону  $[-1;+1]$ . Это позволяет использовать разнотипные показатели в составе одного входного вектора нейронной сети и обеспечивает согласование масштаба данных с областью значений функции активации  $\text{tansig}$ .

Параметры нормализации, рассчитанные на обучающей части данных, далее применялись к валидационным и тестовым наблюдениям без повторной подгонки. Такой порядок обработки необходим для предотвращения утечки информации из тестовой выборки в процесс обучения модели. После получения прогнозных значений в нормализованном пространстве выполнялось обратное преобразование, позволяющее представить прогноз урожайности в исходных единицах измерения – ц/га.

После выполнения указанных процедур был сформирован подготовленный массив данных, пригодный для построения нейросетевой модели. На данном этапе входные признаки, различающиеся по единицам измерения и числовым диапазонам, были приведены к единому интервалу значений; информативные экстремальные наблюдения сохранены в составе выборки, а входные и выходные матрицы подготовлены в формате, необходимом для обучения нейронной сети в среде MATLAB.

#### **2.4 Обоснование архитектуры нейронной сети для прогнозирования урожайности пшеницы**

На основании результатов разведочного анализа и предварительной обработки данных была определена архитектура нейросетевой модели для прогнозирования урожайности яровой пшеницы. В настоящем исследовании задача прогнозирования рассматривается как задача регрессии, поскольку целевая переменная является непрерывной числовой величиной и выражается в центнерах с гектара.

Выбор многослойного персептрона (Multilayer Perceptron, MLP) обусловлен тем, что результаты разведочного анализа позволили предположить наличие нелинейных и пороговых зависимостей между предикторами и урожайностью, что требует применения модели, способной аппроксимировать сложные зависимости.

Теоретическим основанием применения MLP служит теорема об универсальной аппроксимации: в работе К. Хорника, М. Стинчкомба и Х. Уайта показано, что стандартные многослойные сети прямого распространения с нелинейными функциями активации способны аппроксимировать широкий класс функций при достаточном числе скрытых нейронов [46]. Независимое доказательство данного свойства для сигмоидальных активационных функций получено Цыбенко [47].

В качестве архитектурной основы модели использована нейронная сеть прямого распространения сигнала, относящаяся к классу многослойных персептронов. Многослойный персептрон представляет собой искусственную нейронную сеть, состоящую из входного слоя, одного или нескольких скрытых слоев и выходного слоя. В такой архитектуре сигнал распространяется последовательно от входных нейронов к скрытым слоям и далее к выходному нейрону, при этом связи между нейронами соседних слоев имеют настраиваемые весовые коэффициенты. В отличие от рекуррентных сетей, MLP не содержит обратных связей, поэтому относится к сетям прямого распространения сигнала. В настоящем исследовании MLP используется для решения задачи регрессии, поскольку модель преобразует вектор из 16 входных признаков  $X_1$ – $X_{16}$  в одно непрерывное выходное значение  $\hat{Y}$  – прогнозируемую урожайность пшеницы. В такой сети данные последовательно передаются от входного слоя к скрытому слою и далее к выходному слою без обратных связей внутри архитектуры. Согласно документации MathWorks, функция `feedforwardnet` создает сеть прямого распространения с заданным количеством нейронов в скрытых слоях и выбранной функцией обучения; слои сети последовательно соединяются друг с другом, а последний слой формирует выход сети [48].

Входной слой модели включает 16 признаков ( $X_1$ – $X_{16}$ ), характеризующих запасы продуктивной влаги в слоях 0–10 см, 0–50 см и 0–100 см, содержание нитратного азота и подвижного фосфора, количество осадков и температуру воздуха за май–август, а также фитосанитарные показатели. Выходной слой представлен одним нейроном, формирующим прогнозное значение урожайности пшеницы ( $\hat{Y}$ ). Формально нейросетевая модель реализует нелинейное отображение (2.19):

$$\hat{Y} = F(X, W, b), \quad (2.19)$$

где  $X = \{X_1, X_2, \dots, X_{16}\}$  — вектор входных признаков;  
 $W$  – совокупность весовых коэффициентов сети;  
 $b$  – совокупность смещений;

$\hat{Y}$  – прогнозируемая урожайность пшеницы, ц/га.

Обучение нейронной сети состоит в подборе  $W$  и  $b$ , при которых ошибка между фактическими значениями урожайности  $Y$  и прогнозными значениями  $\hat{Y}$  становится минимальной.

В скрытом слое использована функция активации *tansig*, представляющая собой гиперболическую тангенциальную сигмоидальную функцию. Согласно документации MathWorks, функция *tansig* используется как передаточная функция слоя нейронной сети и возвращает значения в диапазоне от  $-1$  до  $+1$  [49]. В математическом виде (2.20):

$$\textit{tansig}(z) = \frac{2}{1 + e^{-2z}} - 1, \quad (2.20)$$

где  $z$  – взвешенная сумма входов нейрона.

Использование нелинейной функции активации в скрытом слое позволяет модели учитывать нелинейные зависимости между входными признаками и урожайностью. В выходном слое применена линейная функция активации *purelin*, что соответствует регрессионному характеру задачи: функция возвращает значение, равное входному сигналу, формируя непрерывное прогнозное значение [30]. Важно разграничить два этапа построения нейросетевой модели. Первый этап связан с выбором числа нейронов скрытого слоя и выполнялся с использованием 5-кратной кросс-валидации (*5-fold cross-validation*) как процедуры сравнения различных конфигураций сети. При такой схеме исходная выборка делится на пять частей, каждая из которых поочередно используется для проверки модели, тогда как оставшиеся части применяются для обучения. Вторым этапом является обучение выбранной архитектуры и выполнялся по схеме разбиения выборки  $70\%/15\%/15\%$ , где данные распределялись на обучающую, валидационную и тестовую подвыборки. Такое разграничение позволяет методически разделить процедуру подбора архитектуры и процедуру итоговой оценки качества модели: 5-кратная кросс-валидация использовалась для выбора числа нейронов скрытого слоя, а схема  $70/15/15$  – для финального обучения и тестовой оценки выбранной модели.

Для определения числа нейронов скрытого слоя был выполнен последовательный перебор значений от 1 до 15. Качество каждой конфигурации оценивалось с применением 5-кратной кросс-валидации. На каждой итерации кросс-валидации одна часть выборки использовалась для проверки модели, а остальные четыре – для ее обучения. Для снижения влияния случайной инициализации весов обучалось несколько сетей с различными начальными условиями. В качестве критериев сравнения конфигураций использовались

среднеквадратическая ошибка на валидационной подвыборке  $val\_MSE$ , корень из среднеквадратической ошибки  $val\_RMSE$  и коэффициент детерминации  $val\_R^2$ . Показатель  $MSE$  характеризует средний квадрат отклонения прогнозных значений от фактических,  $RMSE$  является его корнем и показывает величину ошибки в масштабе анализируемой переменной, а  $R^2$  отражает долю дисперсии целевой переменной, объясняемую моделью. В данном разделе указанные метрики используются для выбора архитектуры сети; их математическое определение и итоговая интерпретация приведены в разделе 2.5.

Результаты перебора числа нейронов скрытого слоя представлены в таблице 2.5 и на рисунке 2.7.

Таблица 2.5 – Зависимость валидационной ошибки от числа нейронов скрытого слоя при 5-кратной кросс-валидации

Нейроны	$val\_MSE$	$val\_RMSE$	$val\_R^2$
1	0,2482	0,4982	-0,032
2	0,2400	0,4899	0,010
3	0,2327	0,4824	0,020
4	0,2591	0,5090	-0,088
5	0,2568	0,5067	-0,104
6	0,2672	0,5169	-0,137
7	0,2690	0,5187	-0,174
8	0,2456	0,4956	-0,098
9	0,3103	0,5571	-0,545
10	0,3060	0,5532	-0,733
11	0,3424	0,5851	-0,790
12	0,3320	0,5762	-0,537
13	0,2918	0,5402	-0,398
14	0,3229	0,5683	-0,654
15	0,3799	0,6164	-0,873

Анализ данных таблицы 2.5 показывает, что наименьшее значение валидационной ошибки получено при использовании трех нейронов скрытого слоя:  $val\_MSE=0,2327$ ,  $val\_RMSE=0,4824$ . Дальнейшее увеличение числа нейронов не приводит к устойчивому повышению качества модели. Напротив, для большинства конфигураций с числом нейронов более трех наблюдается рост ошибки, а значение  $val\_R^2$  становится отрицательным. Отрицательные значения  $R^2$  указывают на то,

что соответствующая конфигурация хуже воспроизводит целевую переменную, чем простая модель, использующая среднее значение целевой переменной в качестве прогноза.

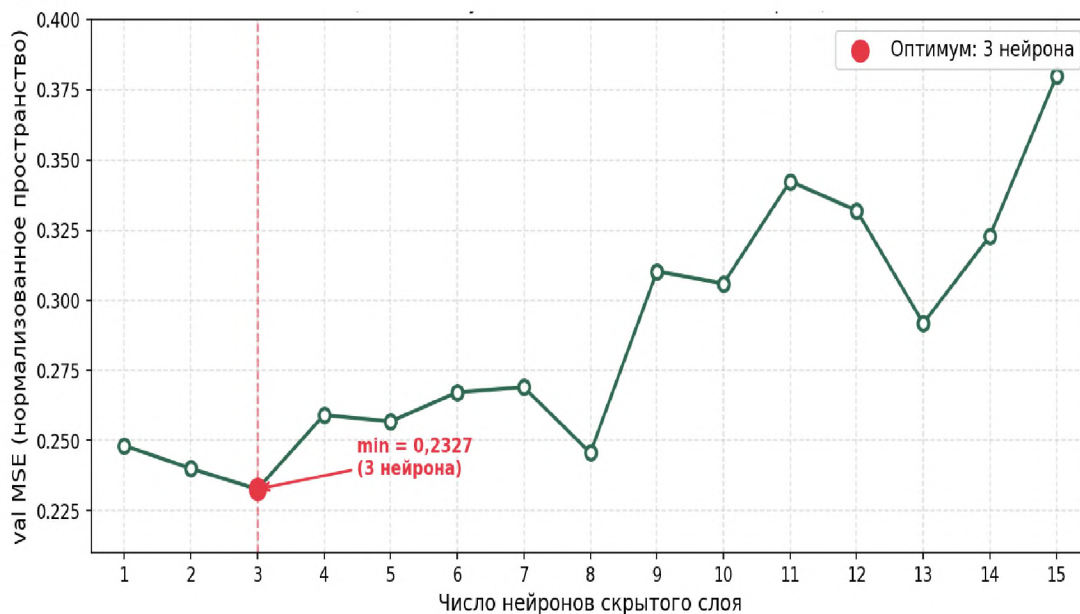


Рисунок 2.7 – Зависимость валидационной ошибки от числа нейронов скрытого слоя при 5-кратной кросс-валидации

Следует учитывать, что на этапе подбора архитектуры обучение выполнялось в поисковом режиме: для каждой конфигурации обучалось несколько сетей с различными начальными условиями без применения финального ансамблирования. В связи с этим при увеличении числа нейронов скрытого слоя признаки переобучения проявляются более заметно: рост количества обучаемых параметров не компенсируется объемом обучающих данных, что приводит к ухудшению валидационных показателей. Представленная на рисунке 2.7 зависимость визуально подтверждает данный результат. На основании проведенного перебора в качестве итоговой архитектуры нейросетевой модели выбрана структура 16–3–1, включающая 16 входных нейронов, 3 нейрона скрытого слоя и 1 выходной нейрон.

Количество обучаемых параметров нейронной сети рассчитывается как сумма весовых коэффициентов и смещений между соседними слоями. Для сети с одним скрытым слоем общее число обучаемых параметров определяется следующим образом:

$$P = (m \cdot h + h) + (h \cdot q + q), \quad (2.21)$$

где  $P$  – общее число обучаемых параметров;

$m$  – число входных признаков;  
 $h$  – число нейронов скрытого слоя;  
 $q$  – число нейронов выходного слоя.

Первое слагаемое ( $m \cdot h + h$ ) соответствует весам и смещениям между входным и скрытым слоями, второе слагаемое ( $h \cdot q + q$ ) – весам и смещениям между скрытым и выходным слоями.

Для выбранной архитектуры 16–3–1 значения параметров составляют:  $m=16$ ,  $h=3$ ,  $q=1$ . Тогда количество обучаемых параметров равно:

$$P = (16 \cdot 3 + 3) + (3 \cdot 1 + 1) = 55. \quad (2.22)$$

При объеме выборки  $N=182$  отношение числа наблюдений к числу обучаемых параметров составляет:

$$\frac{N}{P} = \frac{182}{55} \approx 3,3 \quad (2.23)$$

Данное соотношение использовалось как дополнительная эвристическая проверка соответствия сложности модели объему исходных данных. Оно не является строгим универсальным критерием, однако показывает, что выбранная архитектура имеет умеренное число обучаемых параметров. Для снижения риска переобучения дополнительно использовались ранняя остановка обучения ( $\text{max\_fail}=6$ ) и ансамблирование результатов, при котором итоговый прогноз формировался на основе усреднения выходов лучших сетей.

Для обучения нейронной сети использован алгоритм Левенберга–Марквардта, реализованный в MATLAB функцией `trainlm`. Алгоритм обучения не определяет архитектурный тип сети, а задает способ настройки ее весовых коэффициентов и смещений. Согласно документации MathWorks, функция `trainlm` обновляет веса и смещения сети в соответствии с алгоритмом Левенберга–Марквардта и относится к быстрым алгоритмам обучения для малых и средних нейронных сетей [32]. В работе М. Хагана и М. Менхаджа алгоритм Марквардта был адаптирован для обучения сетей прямого распространения и показал эффективность при решении задач аппроксимации функций [23].

Метод Левенберга–Марквардта относится к численным методам оптимизации нелинейных задач наименьших квадратов и сочетает свойства метода Гаусса–Ньютона и градиентного спуска. В общем виде обновление весовых коэффициентов может быть записано следующим образом:

$$\Delta w = -(J^T J + \mu I)^{-1} J^T e, \quad (2.24)$$

где  $J$  – матрица Якоби, содержащая частные производные ошибок по весовым коэффициентам;

$e$  – вектор ошибок обучающей выборки;

$\mu$  – параметр демпфирования;

$I$  – единичная матрица.

При малых значениях  $\mu$  алгоритм приближается к методу Гаусса–Ньютона, что обеспечивает быструю сходимость в окрестности минимума. При больших значениях  $\mu$  метод ведет себя ближе к градиентному спуску, что повышает устойчивость процесса оптимизации [50]. Такая адаптивность делает алгоритм Левенберга–Марквардта удобным для обучения нейронных сетей в задачах регрессии при ограниченном объеме выборки.

Инициализация весов выполнялась методом Нгуена-Видроу (initnw), который используется для выбора начальных значений весовых коэффициентов и способствует более равномерному распределению активных областей нейронов во входном пространстве [51]. Параметры итоговой нейросетевой модели и ее обучения сведены в таблицу 2.6.

Таблица 2.6 – Параметры итоговой нейросетевой модели.

Параметр	Значение	Обоснование
Архитектура	16–3–1	Минимум val_MSE при переборе 1–15 нейронов скрытого слоя с использованием 5-кратной кросс-валидации
Функция активации скрытого слоя	tansig	Нелинейная функция; диапазон [–1; +1] согласован с нормализацией
Функция активации выходного слоя	purelin (линейная)	Задача регрессии; формирование непрерывного числового выхода
Алгоритм обучения	Левенберга-Марквардта (trainlm)	Настройка весов и смещений сети; высокая скорость сходимости для малых и средних сетей
Нормализация данных	Mapminmax [–1; +1]	Приведение признаков с различными единицами измерения к единому диапазону
Инициализация весов	Нгуена-Видроу (initnw)	Формирование начальных значений весов для более равномерного распределения активных областей нейронов
Разбиение для финального обучения	70% / 15% / 15%	Обучающая, валидационная и тестовая подвыборки

Параметр	Значение	Обоснование
Критерий остановки	ранней $\max\_fail = 6$ эпох	Остановка при отсутствии улучшения $val\_MSE$ ; снижение риска переобучения
Максимум эпох	1 000	Ограничение длительности процесса обучения
Число обучаемых параметров	$55 (16 \times 3 + 3 + 3 \times 1 + 1)$	$N/P = 182/55 \approx 3,3$ ; дополнительная эвристическая проверка соответствия сложности модели объему выборки

Итоговая нейросетевая модель имеет структуру 16–3–1 и включает входной слой из 16 признаков, один скрытый слой с 3 нейронами и выходной слой с одним нейроном. В скрытом слое используется функция активации  $tansig$ , а в выходном слое – линейная функция  $purelin$ . Все соединения между соседними слоями являются полносвязными. Основные параметры модели и процедуры обучения представлены в таблице 2.6, а схема разработанной архитектуры приведена на рисунке 2.8.

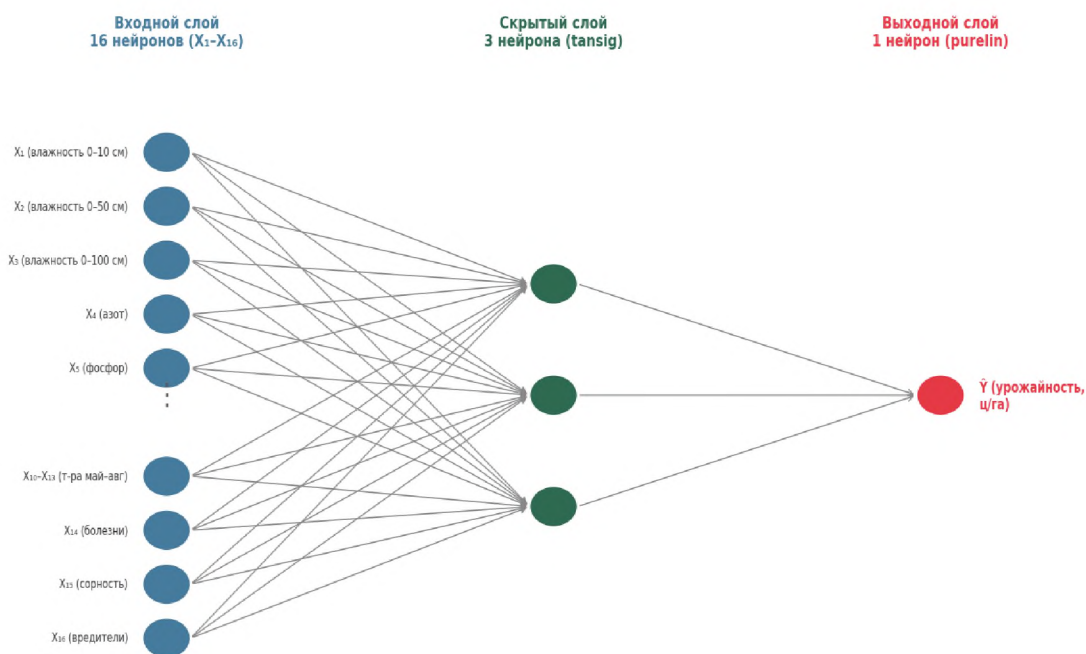


Рисунок 2.8 – Архитектура нейросетевой модели 16–3–1

На рисунке 2.8 представлена структура разработанной нейросетевой модели. Входной слой включает 16 нейронов, соответствующих признакам  $X_1$ – $X_{16}$ . Скрытый слой содержит 3 нейрона с функцией активации  $tansig$ , а выходной слой представлен одним нейроном с линейной функцией активации  $purelin$ ,

формирующим прогнозное значение урожайности  $\hat{Y}$ . Стрелками показаны полностью связанные соединения между соседними слоями.

В разделе 2.4 определена и обоснована архитектура нейросетевой модели прогнозирования урожайности пшеницы. В качестве базовой структуры выбран многослойный персептрон прямого распространения сигнала с архитектурой 16–3–1, содержащий 55 обучаемых параметров. Число нейронов скрытого слоя определено на основе перебора конфигураций и оценки валидационной ошибки методом 5-кратной кросс-валидации. Обучение итоговой модели выполняется с использованием алгоритма Левенберга–Марквардта (trainlm). Оценка качества модели проводится в два этапа: предварительная тестовая оценка по схеме 70% / 15% / 15% и основная оценка методом 5-кратной ансамблевой кросс-валидации, результаты которой рассматриваются в разделе 2.5.

### **2.5 Обучение и оценка качества нейросетевой модели**

Оценка качества нейросетевой модели проводилась в два этапа. На первом этапе итоговая архитектура 16–3–1 обучалась по схеме 70%/15%/15%, где данные разделялись на обучающую, валидационную и тестовую подвыборки. Данный этап использовался для предварительной тестовой оценки выбранной архитектуры на части исходного набора данных

На втором этапе применялась 5-кратная кросс-валидация с ансамблированием. Данный этап рассматривался как основной, поскольку позволял оценить устойчивость модели при различных разбиениях исходной выборки. На каждой итерации кросс-валидации обучалось 50 нейронных сетей с различными начальными условиями. Из них отбирались 10 лучших сетей по значению валидационной ошибки, а итоговый прогноз формировался как среднее значение их выходов. Такой подход позволил снизить влияние случайной инициализации весовых коэффициентов на результат прогнозирования.

Применение ансамблирования повысило среднее значение  $R^2$  по итерациям кросс-валидации с 0,37 для одиночной сети до 0,58 для ансамблевой модели. Прирост на 0,21 свидетельствует о целесообразности применения ансамблевой стратегии для повышения устойчивости прогноза в рамках проведенного вычислительного эксперимента.

Для оценки устойчивости и обобщающей способности модели использовалась 5-кратная кросс-валидация (5-fold cross-validation). Исходный набор данных объемом  $N=182$  случайным образом разделялся на пять непересекающихся частей по 36–37 наблюдений. На каждой итерации одна часть выборки использовалась для проверки модели, а остальные четыре – для обучения ансамбля. Стратификация по районам обеспечивала присутствие наблюдений из каждого из 13 районов Северо-Казахстанской области в проверочной подвыборке.

Для обеспечения воспроизводимости вычислительного эксперимента в MATLAB фиксировалось начальное состояние генератора псевдослучайных чисел с помощью команды `rng(42)`. Это позволяло получать одинаковое разбиение данных

на подвыборки и одинаковые начальные условия при повторном запуске программного кода. Число 42 в данном случае является фиксированным значением начального состояния генератора и не имеет самостоятельного содержательного значения.

Следует отметить ограничение примененной схемы. Случайное разбиение по наблюдениям допускает ситуацию, при которой данные одного и того же года могут одновременно присутствовать в обучающей и проверочной подвыборках, но для разных районов. Методологической альтернативой является схема исключения одного года (leave-one-year-out), при которой все наблюдения за один год полностью исключаются из обучающей выборки и используются только для тестирования модели. Затем процедура повторяется последовательно для каждого года наблюдений. Такая схема позволяет проверить способность модели прогнозировать урожайность для полностью нового года, однако при ограниченном объеме данных она может приводить к повышенной изменчивости оценок качества. Поэтому в настоящем исследовании 5-кратная кросс-валидация рассматривалась как более устойчивая внутренняя оценка качества модели в пределах имеющейся выборки, сформированной по схеме «район–год».

Для оценки качества прогноза использовались четыре показателя: среднеквадратическая ошибка MSE, коэффициент детерминации  $R^2$ , корень из среднеквадратической ошибки RMSE и средняя абсолютная процентная ошибка MAPE. В процессе обучения и валидации алгоритм Левенберга–Марквардта минимизирует среднеквадратическую ошибку MSE, которая определяется следующим образом:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (2.25)$$

где  $y_i$  – фактическое значение урожайности в  $i$ -м наблюдении;  
 $\hat{y}_i$  – прогнозное значение урожайности;  
 $N$  – количество наблюдений.

Коэффициент детерминации  $R^2$  показывает долю дисперсии целевой переменной, объясняемую моделью:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (2.26)$$

где  $\bar{y}$  – среднее значение урожайности по выборке.

Среднеквадратическая ошибка RMSE характеризует абсолютную величину ошибки прогноза в исходных единицах измерения:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (2.27)$$

Средняя абсолютная процентная ошибка MAPE использовалась для оценки относительной ошибки прогноза:

$$MAPE = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2.28)$$

Интерпретация значений MAPE выполнялась в соответствии с классификацией Льюиса [27]: MAPE < 10% соответствует высокоточному прогнозу, 10–20% – точному прогнозу, 20–50% – приемлемому прогнозу, а значения выше 50% – неточному прогнозу.

Сводные результаты двух этапов оценки представлены в таблице 2.7.

Таблица 2.7 – Сводные результаты оценки качества нейросетевой модели 16–3–1

Этап оценки	Схема	R <sup>2</sup>	RMSE, ц/га	MAPE, %	Категория по Льюису [
Предварительная тестовая оценка	70% / 15% / 15%, тестовая часть около 27 наблюдений	0,625	2,180	12,10	Точный прогноз
Основная оценка	5-кратная ансамблевая кросс-валидация, вся выборка	0,582±0,087	2,308	13,39	Точный прогноз

Результаты предварительного этапа по схеме 70% / 15% / 15% показали следующие значения: R<sub>тест</sub>=0,7906, R<sup>2</sup><sub>тест</sub>≈0,625, RMSE<sub>тест</sub>=2,180 ц/га, MAPE<sub>тест</sub>=12,10%. Согласно классификации Льюиса, значение MAPE=12,10% соответствует категории точного прогноза. Следует учитывать, что тестовая подвыборка в данной схеме включает около 27 наблюдений, поэтому такая оценка может зависеть от состава конкретного разбиения. Итоговые результаты основного этапа, выполненного методом 5-кратной ансамблевой кросс-валидации по всей выборке, составили: R<sup>2</sup>=0,582±0,087, RMSE=2,308 ц/га, MAPE=13,39%. По шкале Льюиса данное значение MAPE также относится к категории точного прогноза. Несколько более высокое значение RMSE при 5-кратной кросс-валидации (2,308 против 2,180 ц/га) при одновременно меньшем R<sup>2</sup> (0,582 против 0,625) является закономерным результатом: тестовая подвыборка из ~27 наблюдений по схеме

70/15/15 охватывает лишь часть изменчивости исходной выборки и зависит от конкретного разбиения, тогда как 5-кратная кросс-валидация последовательно проверяет модель на всех 182 наблюдениях и формирует более представительную и консервативную оценку качества. Сводный график сопоставления фактической и прогнозной урожайности на тестовой выборке представлен на рисунке 2.9.

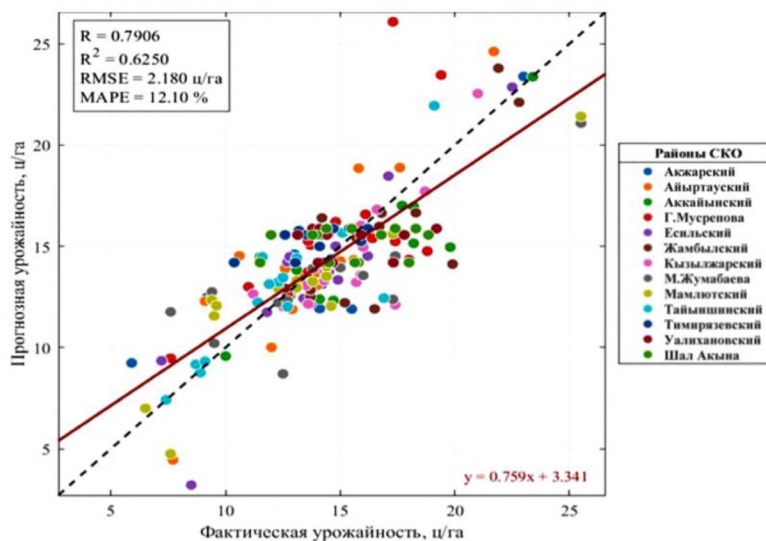


Рисунок 2.9 – Диаграмма рассеяния фактической и прогнозной урожайности нейросетевой модели 16–3–1

На рисунке 2.9 показано сопоставление фактических и прогнозных значений урожайности пшеницы. Пунктирная линия соответствует идеальному прогнозу  $y=x$ , при котором прогнозные значения полностью совпадают с фактическими. Красная линия отражает линейную регрессию прогнозных значений относительно фактических, а цвет точек соответствует отдельным районам Северо-Казахстанской области. Анализ рисунка 2.9 показывает, что основная часть точек сосредоточена вблизи линии идеального прогноза, что свидетельствует о согласованности фактических и расчетных значений урожайности. Наклон линии регрессии прогнозов ( $\beta \approx 0,76$ ) указывает на эффект сглаживания: модель несколько занижает высокие значения урожайности и завышает низкие. Такой эффект характерен для моделей, ориентированных на устойчивое обобщение при наличии выраженной межгодовой изменчивости данных. В качестве базовой линии сравнения была использована множественная линейная регрессия с теми же 16 входными признаками  $X_1-X_{16}$ . Сравнение выполнялось по идентичной схеме 5-кратной кросс-валидации. Результаты сравнения представлены в таблице 2.8.

Таблица 2.8 – Сравнение нейросетевой модели 16–3–1 и множественной линейной регрессии

Модель	R <sup>2</sup>	RMSE, ц/га	MAPE, %	Категория по Льюису
Множественная линейная регрессия, 16 предикторов	0,312	2,840	24,7	Приемлемый прогноз
Нейросетевая модель 16–3–1, ансамбль 10 лучших сетей	0,582	2,308	13,39	Точный прогноз
Превосходство нейросетевой модели	+0,270	–18,7%	–11,3 п.п.	На категорию выше

Анализ таблицы 2.8 показывает, что при сравнении по одинаковой схеме кросс-валидации нейросетевая модель 16–3–1 превосходит множественную линейную регрессию по всем рассматриваемым метрикам. Значение коэффициента детерминации увеличивается на 0,270: с 0,312 для линейной регрессии до 0,582 для нейросетевой модели. Значение RMSE снижается на 18,7%, с 2,840 до 2,308 ц/га. Значение MAPE уменьшается на 11,3 процентного пункта: с 24,7% до 13,39%. Согласно шкале Льюиса, множественная линейная регрессия относится к категории приемлемого прогноза, тогда как нейросетевая модель – к категории точного прогноза. Полученные результаты показывают, что применение нелинейной нейросетевой модели позволяет повысить качество прогнозирования урожайности пшеницы по сравнению с линейной моделью. Это связано с тем, что урожайность формируется под воздействием совокупности почвенно-агрохимических, метеорологических и фитосанитарных факторов, связь между которыми не сводится к простой линейной зависимости. Общая схема двухэтапной процедуры оценки качества представлена на рисунке 2.10.

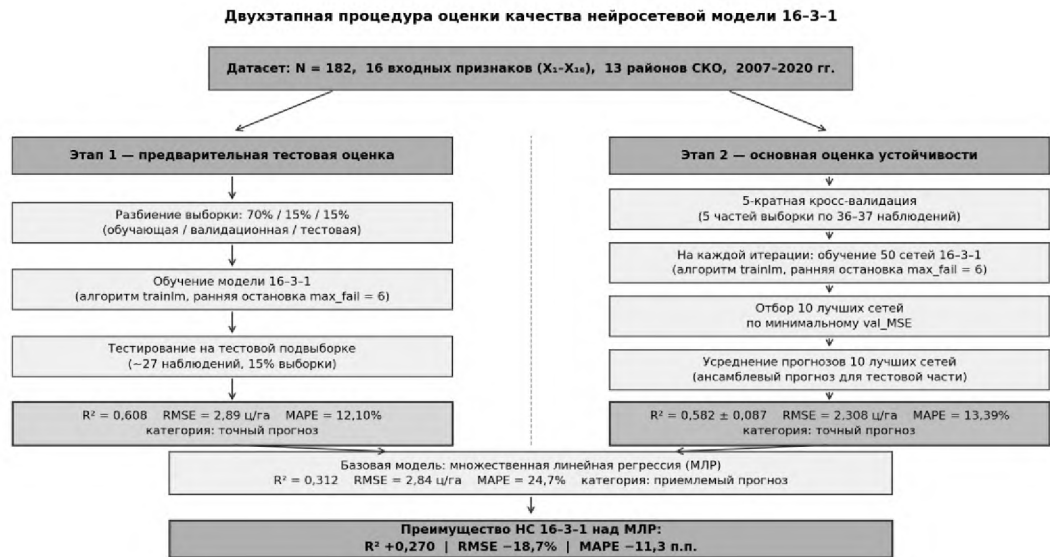


Рисунок 2.10 – Двухэтапная процедура оценки качества нейросетевой модели 16-3-1

На рисунке 2.10 обобщена процедура оценки качества разработанной нейросетевой модели. Первый этап включает предварительное обучение и тестирование по схеме 70%/15%/15%, позволяющее получить первичные тестовые метрики выбранной архитектуры. Вторым этапом является оценка устойчивости модели при изменении состава обучающих и проверочных подвыборок. Дополнительно на данном этапе выполняется сравнение с базовой моделью множественной линейной регрессии.

### Выводы по разделу

Во второй главе диссертационной работы выполнено построение нейросетевой модели для прогнозирования урожайности яровой пшеницы в условиях Северо-Казахстанской области. На первом этапе была сформирована структура исходного набора данных, включающего 13 районов области за период 2007–2020 гг. Общий объем выборки составил 182 наблюдения, каждое из которых соответствует комбинации «район–год». В качестве входных параметров модели использованы 16 признаков, отражающих агрохимические, метеорологические и фитосанитарные условия формирования урожая. Целевой переменной выступила урожайность яровой пшеницы, выраженная в ц/га.

Проведен разведочный анализ данных, позволивший оценить диапазоны изменения переменных, степень их вариабельности, распределение целевой переменной и характер взаимосвязей между признаками. Метод межквартильного размаха IQR использован для первичной диагностики выбросов и проверки качества данных. При этом реальные экстремальные значения урожайности не исключались из выборки, поскольку они отражают фактическую межгодовую

изменчивость агроклиматических условий. Физически невозможные значения метеорологических признаков были идентифицированы как технические ошибки и скорректированы до этапа моделирования.

Описательная статистика показала, что урожайность пшеницы в исследуемой выборке характеризуется умеренной изменчивостью: среднее значение составило 14,52 ц/га, минимальное – 5,9 ц/га, максимальное – 25,5 ц/га. Наиболее высокая вариабельность среди входных признаков выявлена для показателей осадков, что подтверждает существенную межгодовую неустойчивость атмосферного увлажнения в период вегетации. Корреляционный анализ показал, что наиболее выраженная отрицательная связь с урожайностью наблюдается для температуры воздуха в августе, а наиболее заметные положительные связи — для осадков мая и июня. Вместе с тем большинство коэффициентов корреляции имеют слабый или умеренный характер, что подтверждает сложную, многомерную и потенциально нелинейную природу зависимости урожайности от совокупности факторов.

Для подготовки данных к нейросетевому моделированию выполнена нормализация методом `mapminmax`, приводящая значения признаков к диапазону  $[-1;+1]$ . Это позволило согласовать масштабы входных переменных, имеющих различные единицы измерения и числовые диапазоны, с областью значений функции активации `tansig`. После предварительной обработки был сформирован подготовленный массив данных, пригодный для обучения нейронной сети в среде MATLAB R2023b. В качестве архитектурной основы модели выбран многослойный персептрон прямого распространения сигнала. На основе перебора числа нейронов скрытого слоя и оценки валидационной ошибки методом 5-кратной кросс-валидации определена итоговая архитектура 16–3–1, включающая 16 входных нейронов, 3 нейрона скрытого слоя и 1 выходной нейрон. Количество обучаемых параметров данной архитектуры составило 55, что соответствует умеренной сложности модели при имеющемся объеме выборки. В скрытом слое использована нелинейная функция активации `tansig`, в выходном слое – линейная функция `purelin`, соответствующая регрессионному характеру задачи. Обучение нейросетевой модели выполнено с использованием алгоритма Левенберга-Марквардта (`trainlm`). Данный алгоритм применялся для настройки весовых коэффициентов и смещений сети с целью минимизации среднеквадратической ошибки. Для снижения риска переобучения использовалась ранняя остановка обучения, а для повышения устойчивости итогового прогноза применялось ансамблирование результатов.

Оценка качества модели проведена в два этапа. На предварительном этапе использовалась схема разбиения 70% / 15% / 15%, позволившая получить тестовые метрики выбранной архитектуры. На основном этапе применялась 5-кратная ансамблевая кросс-валидация по всей выборке. Итоговые результаты основного этапа составили:  $R^2=0,582\pm 0,087$ ,  $RMSE=2,308$  ц/га,  $MAPE=13,39\%$ . Согласно шкале Льюиса, полученное значение MAPE соответствует категории точного прогноза.

Для оценки эффективности предложенного подхода выполнено сравнение нейросетевой модели с множественной линейной регрессией, построенной на тех же 16 входных признаках. Сравнение по идентичной схеме 5-кратной кросс-валидации показало преимущество нейросетевой модели по всем рассматриваемым метрикам: коэффициент детерминации увеличился с 0,312 до 0,582, значение RMSE снизилось с 2,840 до 2,308 ц/га, а MAPE уменьшилось с 24,7% до 13,39%. Таким образом, нейросетевая модель перешла в категорию точного прогноза, тогда как множественная линейная регрессия относится к категории приемлемого прогноза.

Полученные результаты подтверждают целесообразность применения нейросетевого подхода для прогнозирования урожайности пшеницы в условиях Северо-Казахстанской области. Разработанная модель учитывает комплексное влияние почвенно-агрохимических, метеорологических и фитосанитарных факторов и позволяет описывать нелинейные зависимости, которые не могут быть в полной мере отражены линейной регрессионной моделью. Сформированная во второй главе нейросетевая модель является основой для последующей разработки системы прогнозирования урожайности зерновых культур и ее практического применения в задачах поддержки принятия агротехнологических решений.

### **3 РАЗРАБОТКА И АПРОБАЦИЯ УСТРОЙСТВА СБОРА И ПЕРЕДАЧИ ЛОКАЛЬНЫХ ПОЛЕВЫХ ДАННЫХ ДЛЯ НЕЙРОСЕТЕВОГО ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ**

#### **3.1 Структура устройства сбора и передачи локальных полевых данных**

Результаты второй главы показали, что урожайность пшеницы в условиях Северо-Казахстанской области формируется под влиянием комплекса агрометеорологических, агрохимических и фитосанитарных факторов. Построенная региональная нейросетевая модель позволила описать эти зависимости на уровне районных данных и оценить возможность прогнозирования урожайности по совокупности наиболее значимых входных параметров. Однако региональные данные имеют обобщенный характер и не всегда отражают фактическое состояние конкретного посевного участка. В пределах одного района отдельные поля могут существенно различаться по уровню влагообеспеченности, температурному режиму, химическому составу почвы, применяемым агротехническим мероприятиям и локальным микроклиматическим условиям.

В современных исследованиях по точному земледелию подчеркивается, что IoT-системы и сенсорные сети позволяют получать данные о состоянии почвы и микроклимата с высокой пространственной и временной детализацией. Такие системы применяются для мониторинга влажности почвы, температуры, pH, содержания питательных элементов и других параметров, необходимых для принятия агротехнологических решений и построения прогнозных моделей [53, 54]. Кроме того, интеграция сенсорных данных с методами машинного обучения рассматривается как одно из перспективных направлений развития цифрового земледелия, поскольку позволяет переходить от усредненных наблюдений к анализу условий конкретного поля [55].

В связи с этим для перехода от регионального прогноза к прогнозу на уровне конкретного хозяйства требуется формирование локальной базы данных, основанной на измерениях непосредственно с поля. В рамках настоящего исследования такая задача решалась путем разработки устройства сбора, обработки и передачи агрометеорологических и агрохимических данных. Полевая апробация устройства была выполнена в 2024 г. на участке ТОО «Агрофирма «Мамлютская»: измерения проводились с мая по август восемь раз в сутки, с интервалом 3 часа, с регистрацией температуры воздуха и влажности почвы на глубинах 0,1 м, 0,5 м и 1 м [56]. Полученные локальные данные в дальнейшем могут быть агрегированы по суткам, месяцам или фазам вегетации и использованы как информационная основа для построения локальной модели прогнозирования урожайности.

В результате анализа требований к локальному полевому мониторингу была разработана структурная схема устройства сбора, обработки и передачи агрометеорологических и агрохимических данных. Устройство предназначено для автоматизированной регистрации параметров, характеризующих состояние

посевного участка, и последующей передачи полученной информации для формирования локальной базы данных прогнозирования урожайности.

Структурно система включает два основных модуля: переносной автономный модуль сбора данных и стационарный модуль приема данных. Переносной модуль устанавливается непосредственно на поле и выполняет функции измерения, первичной обработки, отображения, сохранения и передачи данных. В его состав входят датчики влажности почвы, температуры почвы, температуры воздуха и pH почвы, микроконтроллер, часы реального времени, модуль памяти, дисплей, кнопочная панель, GSM/GPRS-модуль связи, антенна и автономная система питания. Стационарный модуль предназначен для приема данных от полевого устройства, их отображения и передачи на персональный компьютер для дальнейшей обработки.

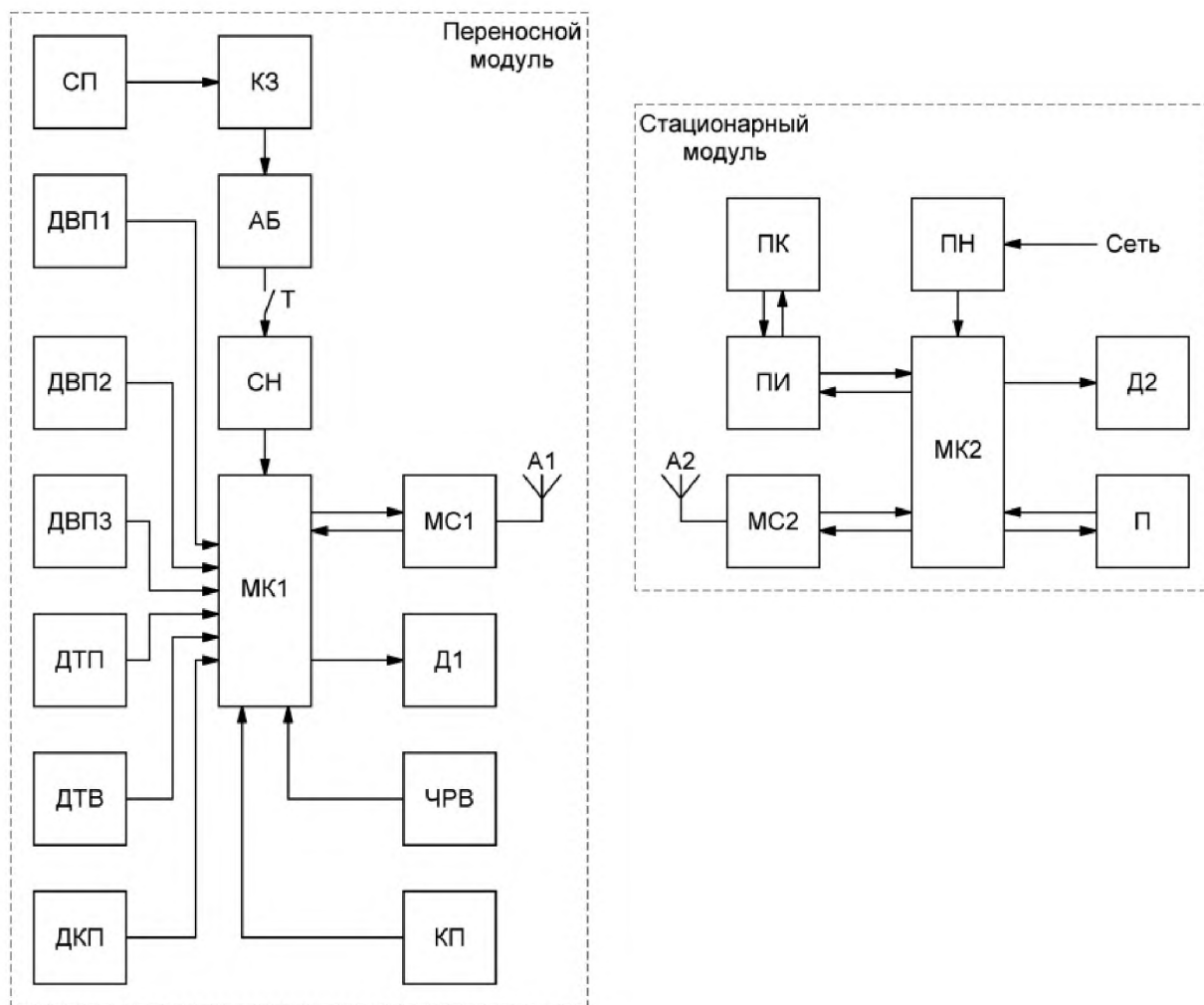


Рисунок 3.1 – Структурная схема устройства сбора, обработки и передачи агрометеорологических данных

На рисунке 3.1 приняты следующие сокращения:

СП – солнечная панель;  
ДВП1 – датчик влажности почвы 1 переносного модуля;  
ДВП2 – датчик влажности почвы 2 переносного модуля;  
ДВП3 – датчик влажности почвы 3 переносного модуля;  
ДТП – датчик температуры почвы переносного модуля;  
ДТВ – датчик температуры воздуха переносного модуля;  
ДКП – датчик кислотности почвы переносного модуля;  
КЗ – контроллер заряда переносного модуля;  
АБ – аккумуляторная батарея переносного модуля;  
СН – стабилизатор напряжения переносного модуля;  
МК1 – микроконтроллер переносного модуля;  
МС1 – модуль связи 1 переносного модуля;  
Д1 – дисплей 1 переносного модуля;  
ЧРВ – часы реального времени переносного модуля;  
КП – кнопочная панель переносного модуля;  
А1 – антенна 1 переносного модуля;  
ПК – персональный компьютер;  
ПИ – преобразователь интерфейса стационарного модуля;  
ПН – преобразователь напряжения стационарного модуля;  
МС2 – модуль связи 2 стационарного модуля;  
МК2 – микроконтроллер 2 стационарного модуля;  
Д2 – дисплей 2 стационарного модуля;  
П – модуль памяти стационарного модуля;  
А2 – Антенна 2 стационарного модуля.

Функционирование переносного модуля основано на последовательном выполнении операций измерения, первичной обработки, сохранения и передачи данных. Измерительные датчики регистрируют параметры почвенно-микrokлиматической среды, после чего сигналы поступают на микроконтроллер. Микроконтроллер выполняет роль центрального управляющего элемента: обеспечивает опрос датчиков, преобразование и первичную обработку входных сигналов, присваивает измерениям временную метку, сохраняет результаты в модуле памяти и формирует пакет данных для передачи на стационарное устройство. Датчики влажности почвы подключаются к аналоговым входам микроконтроллера, поскольку формируют выходной сигнал в виде изменяющегося напряжения. Цифровые элементы устройства, включая датчик температуры почвы, часы реального времени и дисплейный модуль, взаимодействуют с микроконтроллером через соответствующие интерфейсы. После первичной обработки микроконтроллер формирует пакет данных, который передается на стационарное принимающее устройство через GSM/GPRS-модуль связи. Благодаря этому обеспечивается регулярное накопление локальных полевых измерений, отражающих фактическое состояние посевного участка и используемых далее при подготовке базы данных для нейросетевого прогнозирования урожайности.

Автономность переносного модуля обеспечивается солнечной фотоэлектрической панелью, аккумуляторной батареей, контроллером заряда и стабилизаторами напряжения. Такое решение обусловлено условиями эксплуатации устройства в открытом поле, где отсутствует возможность постоянного подключения к электрической сети.

Стационарный модуль выполняет прием данных от переносного устройства, их отображение на дисплее и передачу на персональный компьютер. Далее полученные данные могут быть структурированы в табличном виде, проверены на корректность, агрегированы по заданным временным интервалам и использованы при формировании локальной базы наблюдений. Таким образом, разработанная структурная схема обеспечивает полный цикл работы с полевыми данными: от измерения параметров непосредственно на участке до их подготовки для последующего интеллектуального анализа. Основные компоненты устройства и их назначение представлены в таблице 3.1.

Таблица 3.1 – Основные компоненты устройства сбора, обработки и передачи данных

Компонент устройства	Техническое наименование/ характеристика	Количество (шт)	Функциональное назначение
Микроконтроллер	AVR ATmega328, 8-битный микроконтроллер, корпус TQFP-32, тактовая частота до 20 МГц, 10-битный АЦП	1	Управляющий элемент устройства; опрос датчиков, обработка сигналов, запись и передача данных, Преобразование аналоговых сигналов датчиков влажности почвы и кислотности в цифровой вид
Плата расширения	Собственная разработка	1	Для расширения аппаратной платформы для реализации схемы управления устройством
Датчик температуры воздуха	AM2302, цифровой датчик температуры и влажности, питание 3,3–5,5 В, диапазон температуры от –40 до +80 °С	1	Измерение температуры воздуха на исследуемом участке поля
Датчик влажности почвы	CSMSv2.0, емкостный датчик влажности, питание 3,3–5,5 В, аналоговый выход до 3 В	3	Измерение влажности почвы
Датчик температуры почвы	DS18B20, цифровой термометр, герметичный корпус, интерфейс 1-Wire,	1	Измерение температуры почвы

	диапазон от -55 до +125 °С		
Часы реального времени	DS3231 RTC, интерфейс I2C	1	Формирование временной метки при регистрации данных
Модуль памяти	AT24C256, SD-карта/карта памяти	1	Локальная запись данных с датчиков в файл
Антенна	Антенна GSM-модуля	1	Передача и прием беспроводного сигнала
Преобразователь интерфейса	CH340G, USB-последовательный интерфейс	1	Подключение устройства к компьютеру
Солнечная панель	Солнечная фотоэлектрическая панель мощностью 25 Вт, номинальное напряжение 12 В	1	Автономное питание переносного модуля в полевых условиях
Контроллер заряда	HX-3S-01, для литиевых элементов 11,1V 6A BMS 3S HX-3S-01	1	Управление зарядом аккумуляторной батареи от солнечной панели
Аккумуляторная батарея	Два литиевых аккумулятора 18650, номинальное напряжение 3,7 В, емкость 3200 мА·ч	2	Накопление энергии и питание устройства при отсутствии солнечного излучения
Дисплей	LCD WH2004A, символьный ЖК-дисплей 20×4 с I2C-конвертером PCF8574T	1	Отображение статуса устройства и текущих данных с датчиков
Модуль беспроводной приемопередачи	GSM модуль SIM900A	1	Передача собранных данных на серверное устройство

После определения состава основных компонентов была выполнена компоновка электрической принципиальной схемы устройства. Принципиальная схема необходима для объединения сенсорного блока, микроконтроллера, модуля связи, системы питания, индикации и хранения данных в единую аппаратную систему. При ее разработке учитывалась необходимость одновременной работы с аналоговыми и цифровыми датчиками, автономного питания, локального сохранения информации и беспроводной передачи данных на стационарный модуль. Центральным элементом схемы является микроконтроллер ATmega328, который обеспечивает управление всеми функциональными узлами устройства. Датчики влажности почвы подключаются к аналоговым входам микроконтроллера, так как формируют выходной сигнал в виде изменяющегося напряжения. Далее этот сигнал преобразуется встроенным аналого-цифровым преобразователем и используется для расчета значения влажности почвы. Датчик температуры почвы DS18B20 подключается по интерфейсу 1-Wire, что позволяет передавать цифровые значения температуры с минимальным количеством соединительных линий. Часы

реального времени DS3231 и дисплейный модуль подключаются по шине I2C, а GSM/GPRS-модуль SIM900A взаимодействует с микроконтроллером через интерфейс UART. Отдельное место в схеме занимает система питания. Поскольку устройство предназначено для эксплуатации в полевых условиях, питание переносного модуля реализовано от аккумуляторных батарей, заряд которых осуществляется от солнечной панели. Для обеспечения стабильной работы электронных компонентов используются стабилизаторы напряжения 5 В и 3,3 В. Такое решение позволяет согласовать требования различных узлов устройства по питанию и обеспечить устойчивую работу микроконтроллера, датчиков, дисплея и модуля связи. Передача данных реализуется через GSM/GPRS-модуль SIM900A. После считывания и обработки измерений микроконтроллер формирует пакет данных, который передается на стационарное устройство. Для контроля работы устройства предусмотрен жидкокристаллический дисплей, на который выводятся основные параметры и служебные сообщения. Подключение к персональному компьютеру обеспечивается через преобразователь USB-UART, что позволяет выполнять настройку, проверку и дальнейшую обработку данных.

Электрическая принципиальная схема обеспечивает аппаратную реализацию всех функций, заложенных в структурной схеме: измерение параметров почвенно-микrokлиматической среды, первичную обработку данных, присвоение временной метки, локальное сохранение, отображение текущей информации и беспроводную передачу результатов измерений. Принципиальная схема устройства сбора, обработки и передачи агрометеорологических данных представлена на рисунке 3.2.

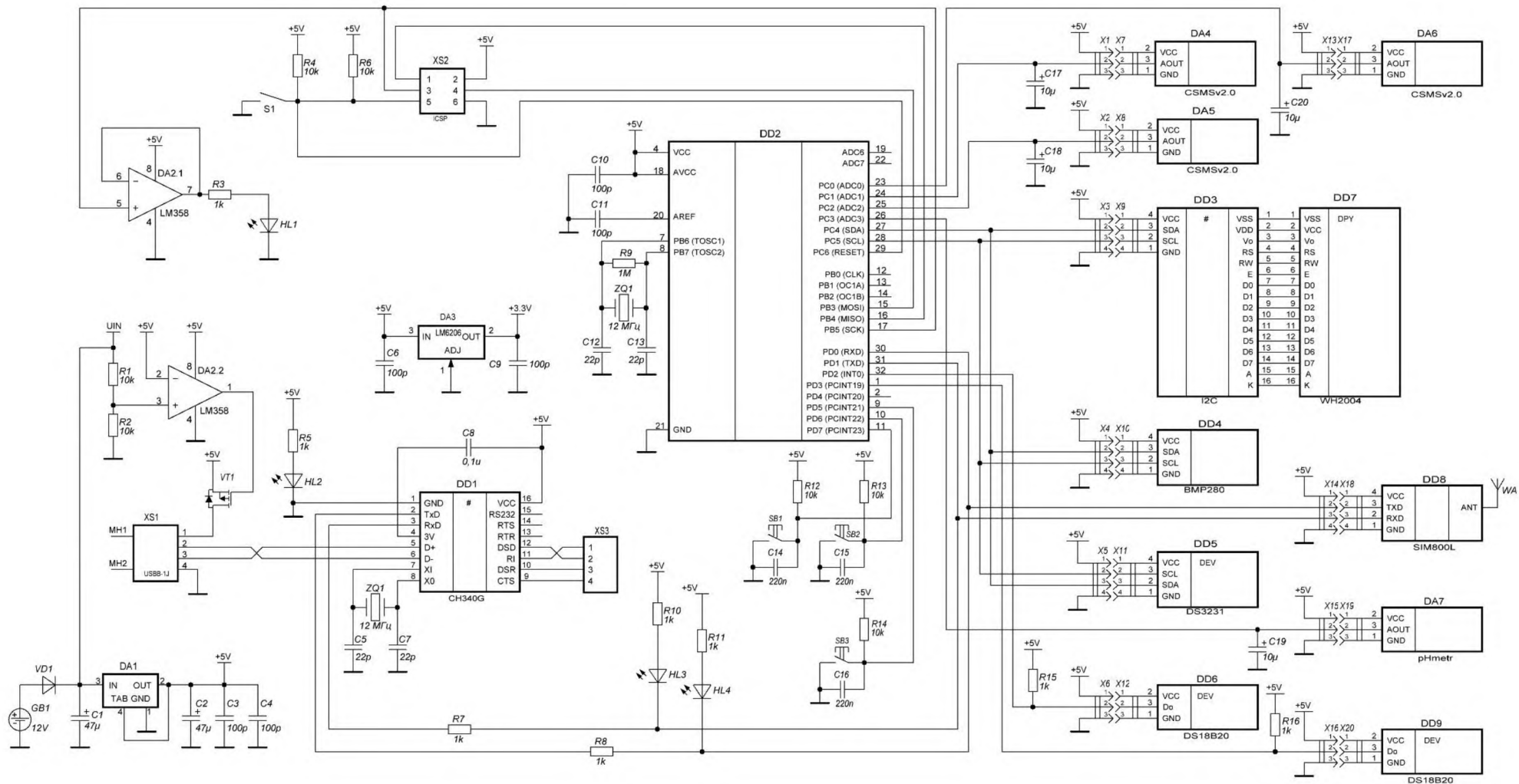


Рисунок 3.2 – Электрическая принципиальная схема устройства сбора, обработки и передачи агрометеорологических данных

На электрической принципиальной схеме показано соединение основных функциональных узлов устройства: сенсорного блока, микроконтроллера, модуля связи, системы питания, индикации, часов реального времени и интерфейса подключения к персональному компьютеру. Центральным элементом схемы является микроконтроллер ATmega328, который выполняет функции управления устройством, опроса датчиков, первичной обработки измерений, формирования временной метки и подготовки данных к сохранению и передаче. Аналоговые сигналы от датчиков влажности почвы поступают на входы АЦП микроконтроллера. Это связано с тем, что емкостные датчики влажности формируют выходной сигнал в виде изменяющегося напряжения, зависящего от состояния почвенной среды. После преобразования сигналов в цифровую форму микроконтроллер использует полученные значения для дальнейшей обработки и записи в память. Цифровой датчик температуры почвы DS18B20 подключается по интерфейсу 1-Wire, часы реального времени DS3231 и дисплейный модуль через PCF8574 – по шине I2C, а GSM/GPRS-модуль SIM900A взаимодействует с микроконтроллером по интерфейсу UART. В схеме также предусмотрена автономная система питания. Источником энергии является аккумуляторная батарея, заряд которой осуществляется от солнечной фотоэлектрической панели через контроллер заряда. Для питания электронных компонентов используются стабилизаторы напряжения, формирующие необходимые уровни 5 В и 3,3 В. Такое решение обеспечивает согласование требований различных узлов устройства и позволяет использовать переносной модуль в полевых условиях без подключения к стационарной электрической сети.

Передача данных осуществляется через GSM/GPRS-модуль SIM900A. После считывания и первичной обработки измерений микроконтроллер формирует пакет данных и передает его на стационарное принимающее устройство. Для контроля работы предусмотрен ЖКИ, на который выводятся текущие параметры и служебные сообщения. Подключение к персональному компьютеру реализовано через преобразователь USB-UART на базе CH340G, что позволяет выполнять настройку устройства и передачу данных для дальнейшей обработки.

### **3.2 Аппаратная и конструктивная реализация устройства**

Проектирование печатной платы выполнялось в программной среде EasyEDA. Данная среда позволяет выполнить построение электрической схемы, переход к проектированию печатной платы, размещение компонентов и трассировку соединений [57]. Использование EasyEDA позволило заранее определить расположение основных узлов устройства и уменьшить вероятность ошибок при последующей сборке. Печатная плата разработана как основа для объединения измерительной, управляющей, коммуникационной и питающей частей устройства. На ней предусмотрено размещение микроконтроллера, разъемов для подключения датчиков влажности почвы, температуры почвы, температуры воздуха и рН почвы, модуля связи, часов реального времени, дисплейного интерфейса, элементов питания и стабилизаторов напряжения.

Такое конструктивное решение позволяет уменьшить количество внешних соединений и повысить надежность работы устройства в условиях полевой эксплуатации.

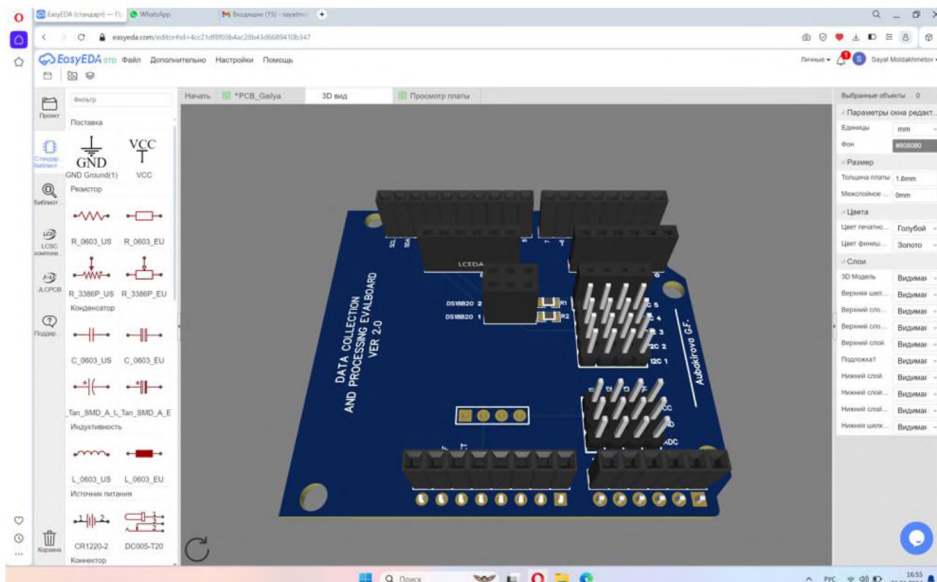
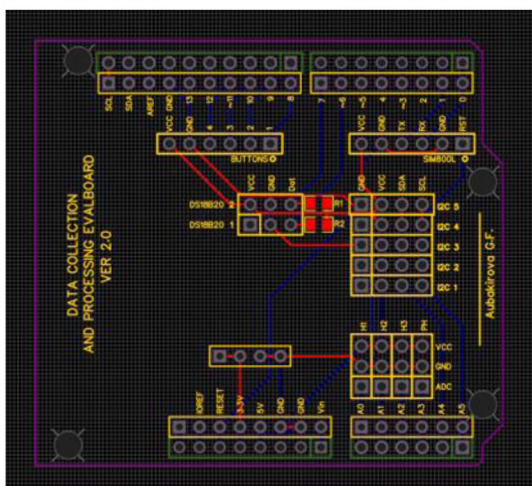
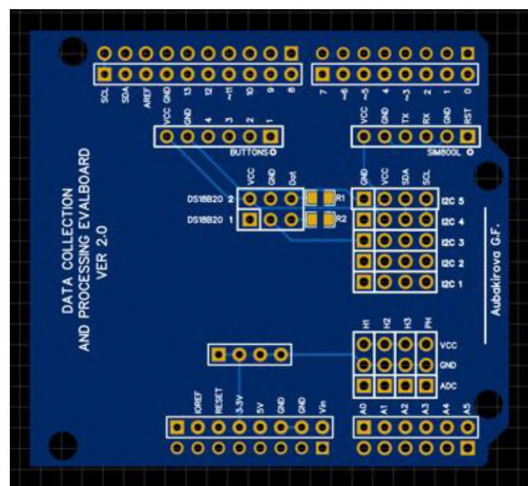


Рисунок 3.3 – Разработка печатной платы устройства в программной среде EasyEDA

Для предварительной оценки компоновки элементов была сформирована 3D-модель печатной платы. Она позволяет визуально проверить размещение компонентов, разъемов и интерфейсных элементов до этапа окончательной сборки. Это особенно важно для автономного полевого устройства, поскольку электронные компоненты должны быть размещены внутри корпуса с учетом подключения внешних датчиков, солнечной панели, антенны и органов управления.



а)



б)

Рисунок 3.4 – 3D-модель печатной платы устройства

а) чертеж печатной платы б) лицевая сторона

После проектирования печатной платы была выполнена сборка опытного образца устройства. Конструктивно устройство размещено в корпусе, предназначенном для защиты электронных компонентов от механических воздействий и внешних условий эксплуатации. В качестве материала корпуса использовано оргстекло, или полиметилметакрилат, толщиной 2,5 мм. Корпус устройства выполнен в виде параллелепипеда размером 100×120×150 мм. Внутри корпуса размещаются печатная плата, элементы питания, соединительные провода и интерфейсные элементы. Автономность устройства обеспечивается за счет сочетания солнечной панели, аккумуляторной батареи, контроллера заряда и стабилизаторов напряжения. Такое решение позволяет использовать устройство в поле без подключения к стационарной электрической сети. При установке устройство размещается на мачте, а солнечная панель ориентируется в сторону наибольшего поступления солнечного излучения. Разработанное устройство предназначено для сбора и передачи полевых данных, необходимых для последующего прогнозирования урожайности зерновых культур [58].

### **3.3 Полевая апробация устройства и анализ локальных измерений**

После сборки опытного образца была проведена полевая апробация разработанного устройства. Испытания выполнялись в 2024 г. на посевном участке ТОО «Агрофирма «Мамлютская» в Мамлютском районе Северо-Казахстанской области. Целью апробации являлась проверка работоспособности устройства в реальных полевых условиях, а также оценка пригодности получаемых данных для последующего формирования локальной базы прогнозирования урожайности.

Устройство было установлено непосредственно на посевном участке. Такое размещение позволило фиксировать параметры не по усредненным районным данным, а в условиях конкретного поля. Измерения выполнялись с мая по август 2024 г. восемь раз в сутки, то есть с интервалом 3 часа. В ходе наблюдений регистрировались температура воздуха и влажность почвы на глубинах 0,1 м, 0,5 м и 1 м. Такой режим измерений позволил получить временные ряды с высокой детализацией и проследить суточную и сезонную динамику почвенно-микrokлиматических условий [59].

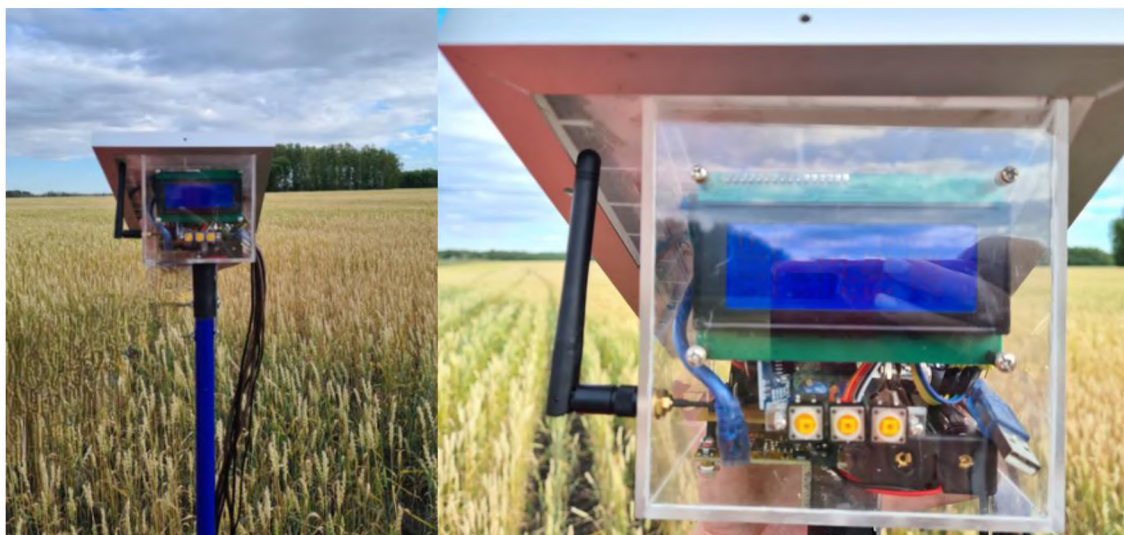


Рисунок 3.5 – Установка устройства сбора полевых данных на посевном участке ТОО «Агрофирма «Мамлютская»

Полученные данные были экспортированы в табличный формат и прошли первичную подготовку. На данном этапе выполнялась проверка временных меток, контроль наличия дублирующихся записей, визуальный просмотр временных рядов и формирование агрегированных показателей. Поскольку измерения выполнялись каждые 3 часа, исходные данные были пригодны для расчета суточных, месячных и сезонных характеристик: средних, минимальных, максимальных значений и амплитуды изменения параметров. В дальнейшем такие агрегированные показатели могут использоваться как входные признаки локальной нейросетевой модели.

Для проверки корректности измерений была выполнена сопоставительная оценка температуры воздуха, полученной с устройства, с официальными метеорологическими данными. На рисунке 3.6 представлено сравнение временных рядов температуры воздуха за май 2024 г. по данным разработанного устройства и Казгидромета.

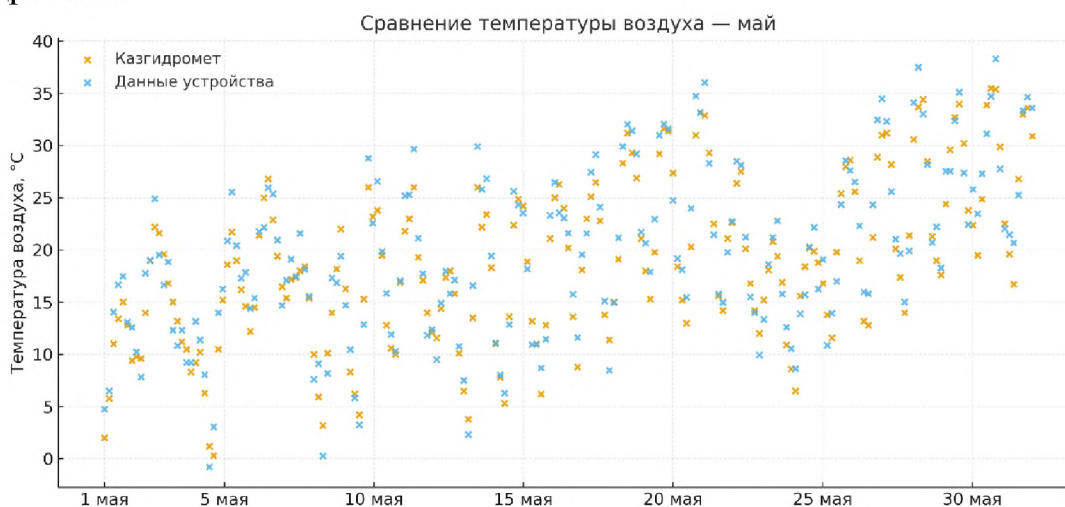


Рисунок 3.6 – Сравнение температуры воздуха по данным устройства и Казгидромета, май 2024 г.

На рисунке 3.6 представлено сравнение температуры воздуха за май 2024 г. по данным Казгидромета и разработанного устройства. Анализ графика показывает, что локальные измерения в целом повторяют динамику официального температурного ряда: периоды повышения и понижения температуры совпадают по времени, а максимальные и минимальные значения наблюдаются в близкие даты. Это свидетельствует о согласованности данных устройства с региональными метеорологическими наблюдениями. При этом между отдельными значениями наблюдаются расхождения, что объясняется различием пространственного уровня измерений. Данные Казгидромета характеризуют температуру воздуха на метеорологической станции, тогда как разработанное устройство фиксирует температуру непосредственно на посевном участке. Поэтому локальные измерения могут отражать микроклиматические особенности конкретного поля, которые не всегда проявляются в усредненных региональных данных. Количественная оценка также подтверждает согласованность двух рядов: коэффициент корреляции Пирсона составил 0,97, среднеквадратическая ошибка – 1,3 °С, средняя абсолютная ошибка – 1,0 °С. Полученные результаты подтверждают пригодность разработанного устройства для локального мониторинга температуры воздуха и последующего формирования признаков для модели прогнозирования урожайности.

Сравнение локальных измерений температуры воздуха с официальными метеорологическими данными подтвердило, что разработанное устройство корректно отражает общую динамику микроклиматических условий на исследуемом участке. Однако для оценки условий формирования урожайности одной температуры воздуха недостаточно, поскольку водный режим почвы является одним из ключевых факторов, определяющих развитие растений в течение вегетационного периода. В связи с этим следующим этапом анализа стало рассмотрение данных о влажности почвы, полученных с датчиков, установленных на разных глубинах. Такой подход позволяет оценить не только поверхностное увлажнение, но и распределение влаги по почвенному профилю, что особенно важно для характеристики условий корневого питания растений.

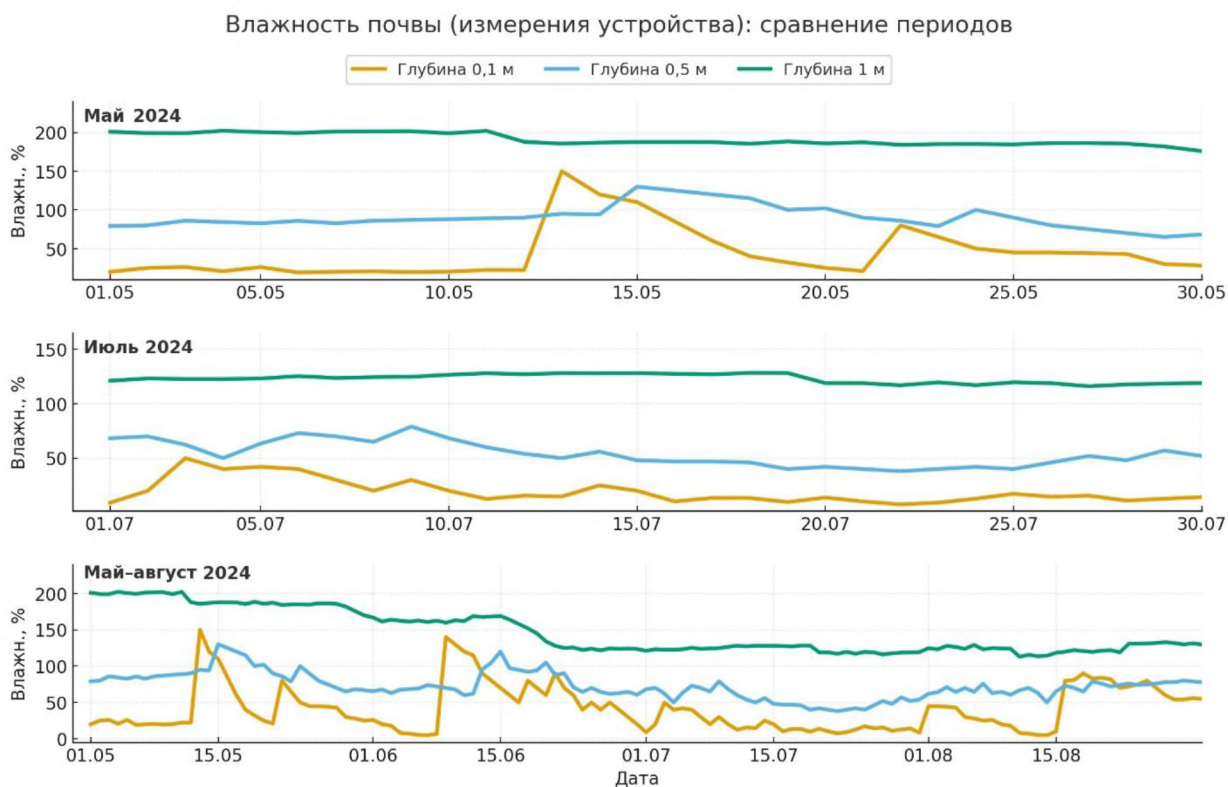


Рисунок 3.7 – Динамика влажности почвы на глубинах 0,1 м, 0,5 м и 1 м по данным устройства

На рисунке 3.7 представлена динамика влажности почвы, измеренной разработанным устройством на трех глубинах: 0,1 м, 0,5 м и 1 м. Для анализа приведены отдельные временные ряды за май, июль, а также обобщенный период май-август. Такое представление позволяет оценить как краткосрочные изменения влажности, так и сезонную динамику водного режима почвы в течение вегетационного периода. Анализ графиков показывает выраженную глубинную дифференциацию влажности почвы. Поверхностный слой на глубине 0,1 м характеризуется наибольшей изменчивостью: значения влажности резко возрастают после поступления влаги и сравнительно быстро снижаются под влиянием испарения, прогрева поверхности и суточных колебаний температуры воздуха. Такая динамика указывает на высокую чувствительность верхнего горизонта к внешним метеорологическим условиям. В то же время данные только с глубины 0,1 м не могут полностью характеризовать влагообеспеченность растений, поскольку этот слой быстро реагирует на кратковременные погодные изменения.

На глубине 0,5 м изменения влажности имеют более сглаженный характер. Это свидетельствует о меньшей чувствительности среднего почвенного горизонта к кратковременным колебаниям внешней среды. Данный слой отражает более устойчивое состояние влаги в почве и может рассматриваться как важный показатель условий корневого питания растений. На глубине 1 м влажность изменяется наиболее плавно и сохраняет сравнительно стабильные значения на протяжении рассматриваемого периода. Такая инерционность глубокого горизонта объясняется замедленной реакцией почвы на осадки, испарение и температурные изменения. Сравнение майского и июльского периодов показывает, что в летний

период наблюдается снижение влажности в верхнем и среднем слоях почвы. Это может быть связано с ростом температуры воздуха, усилением испарения и потреблением влаги растениями в активные фазы вегетации. Обобщенный график за май–август подтверждает сезонную тенденцию изменения водного режима: поверхностный слой демонстрирует резкие колебания, средний слой изменяется более равномерно, а глубокий горизонт сохраняет наибольшую устойчивость.

Полученные результаты подтверждают обоснованность размещения датчиков влажности почвы на нескольких глубинах. Послойное измерение влажности почвы позволяет не только фиксировать текущее состояние верхнего слоя, но и оценивать распределение влаги по почвенному профилю. Для задач прогнозирования урожайности это имеет существенное значение, поскольку формирование урожая зависит не только от кратковременной влажности поверхности, но и от доступности влаги в более глубоких горизонтах, связанных с развитием корневой системы растений.

Полученные данные сенсорного мониторинга представляют собой временные ряды с высокой периодичностью измерений. Поскольку целевая переменная нейросетевой модели – урожайность – определяется по итогам вегетационного периода, трехчасовые измерения не могут рассматриваться как самостоятельные строки обучающей выборки. Для дальнейшего использования в прогнозной модели они предварительно агрегируются до суточных, месячных или сезонных признаков. В частности, по каждому измеряемому параметру могут быть рассчитаны средние, минимальные и максимальные значения, амплитуда изменения, а также характеристики динамики за отдельные периоды вегетации. Такой подход позволяет сохранить информативность исходных временных рядов и привести данные к формату, сопоставимому с сезонным значением урожайности.

### **3.4 Алгоритм преобразования и структурирования полевых данных**

Данные, полученные с разработанного устройства, были преобразованы от первичных сигналов датчиков до структурированной табличной записи, пригодной для дальнейшей обработки и использования в нейросетевой модели прогнозирования урожайности. Измерительные каналы устройства имеют различный тип выходного сигнала: датчики влажности почвы формируют аналоговый сигнал, который требует аналого-цифрового преобразования, а цифровые датчики температуры передают значения в микроконтроллер через соответствующие интерфейсы. Поэтому обработка данных выполнялась с учетом типа измерительного канала.

Необходимость предварительной калибровки емкостных датчиков влажности почвы подтверждается современными исследованиями в области точного земледелия. В научных работах подчеркивается, что выходной сигнал таких датчиков зависит не только от влажности, но и от типа почвы, условий контакта, электропроводности, а также индивидуальных особенностей датчика, [60–64]. Поэтому перед использованием данных в локальной базе прогнозирования была выполнена калибровка датчиков влажности почвы. Такой порядок обработки соответствует подходам, применяемым в системах сенсорного мониторинга почвы: аналоговый сигнал емкостного датчика, сначала преобразуется микроконтроллером в цифровой код, затем полученное значение пересчитывается по калибровочной

зависимости в показатель влажности почвы [65]. В разработанном устройстве преобразование выполняется встроенным 10-битным АЦП микроконтроллера, что позволяет представить входное напряжение в виде цифрового значения в диапазоне от 0 до 1023.

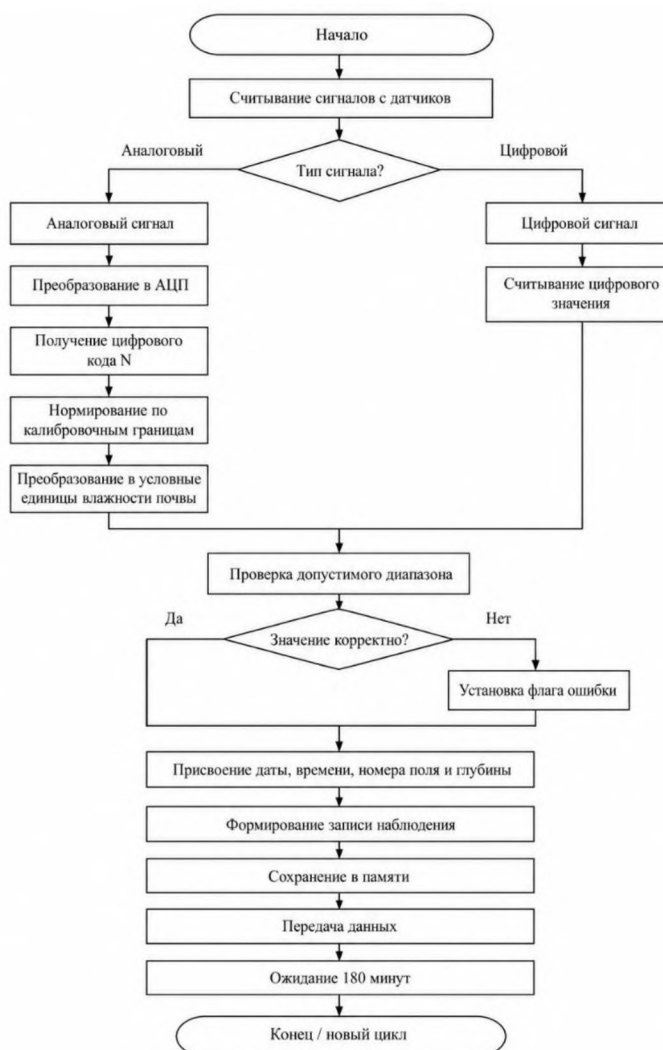


Рисунок 3.8 – Блок-схема алгоритма преобразования и структурирования данных, получаемых с устройства сбора и передачи данных

На рисунке 3.8 представлен алгоритм преобразования данных сенсорного мониторинга в табличную запись. Алгоритм описывает последовательную обработку измерительной информации, начиная со считывания сигналов с датчиков и заканчивая формированием записи, пригодной для последующего использования в локальной базе данных. Аналоговые сигналы датчиков влажности почвы преобразуются встроенным АЦП микроконтроллера в цифровой код, после чего пересчитываются по калибровочной зависимости в условные единицы влажности почвы. Цифровые значения температуры воздуха и температуры почвы добавляются к общей записи после проверки допустимых диапазонов. Каждому измерению присваиваются дата, время, номер поля и глубина установки датчика, после чего запись сохраняется в памяти устройства и передается на стационарный модуль по GSM/GPRS-каналу.

Сигнал с емкостного датчика влажности почвы поступал на аналоговый вход микроконтроллера в виде напряжения  $U_{in}$ , величина которого изменялась в зависимости от состояния почвенной среды. Встроенный 10-битный аналого-цифровой преобразователь микроконтроллера преобразовывал аналоговый сигнал в цифровой код  $N$ . Поскольку разрядность АЦП составляет  $n=10$ , количество уровней квантования равно:

$$2^{10} = 1024, \quad (3.1)$$

При этом цифровой код  $N$  принимает целочисленные значения в диапазоне от 0 до 1023. Следовательно, значение 1024 характеризует количество возможных уровней представления сигнала, а 1023 является максимальным значением цифрового кода.

Шаг квантования АЦП определяется выражением:

$$\Delta U = \frac{U_{ref}}{2^n}, \quad (3.2)$$

где  $\Delta U$  – шаг квантования, В  
 $U_{ref}$  – опорное напряжение АЦП, В;  
 $n$  – разрядность АЦП.

При  $U_{ref} = 5$  В и  $n=10$ :

$$\Delta U = \frac{5}{1024} \approx 0.00488 \text{ В} \quad (3.3)$$

Таким образом, изменению входного напряжения примерно на 4,88 мВ соответствует изменение цифрового кода АЦП на одну единицу. Преобразование входного напряжения  $U_{in}$  в цифровой код может быть представлено следующим выражением:

$$N = \left[ \frac{U_{in}}{U_{ref}} \cdot 2^n \right], \quad (3.4)$$

где  $N$  – цифровой код АЦП;  
 $U_{in}$  – входное напряжение с датчика, В;  
 $U_{ref}$  – опорное напряжение АЦП, В;  
 $n$  – разрядность АЦП.

При этом значение  $N$  ограничивается диапазоном:

$$0 \leq N \leq 1023 \quad (3.5)$$

Полученный цифровой код не использовался в базе данных напрямую. На следующем этапе выполнялся калибровочный пересчет значения АЦП в условный показатель влажности почвы. В общем виде такая зависимость может быть представлена линейной функцией:

$$W_{cond} = aN + b \quad (3.6)$$

где  $W_{cond}$  – условный показатель влажности почвы;  
 $N$  – цифровой код АЦП;  
 $a, b$  – калибровочные коэффициенты, определяемые по результатам настройки датчика.

При использовании двух контрольных состояний датчика коэффициенты линейной зависимости определяются следующим образом:

$$a = \frac{W_2 - W_1}{N_2 - N_1}, \quad (3.7)$$

$$b = W_1 - aN_1, \quad (3.8)$$

где  $N_1, N_2$  – цифровые значения АЦП, полученные при двух контрольных состояниях почвы;

$W_1, W_2$  – соответствующие им значения влажности почвы в условных единицах.

В рамках данной работы такой пересчет использовался для унификации показаний датчиков влажности, установленных на глубинах 0,1 м, 0,5 м и 1 м. При этом показатель  $W_{cond}$  не является прямым измерением запаса продуктивной влаги в агрономическом смысле. Он представляет собой условный сенсорный показатель влажности почвы, рассчитанный на основе сигнала датчика после аналого-цифрового преобразования и калибровочного пересчета.

После преобразования и проверки каждому измерению присваивались дата, время, номер поля и глубина установки датчика. В результате формировалась структурированная запись наблюдения:

$$(ID_{устр}, ID_{поля}, t, W_{0.1}, W_{0.5}, W_{1.0}, T_{air}, T_{soil}, pH, flag), \quad (3.9)$$

где  $ID_{устр}$  – идентификатор устройства;

$ID_{поля}$  – идентификатор поля;

$t$  – временная метка;

$W_{0.1}, W_{0.5}, W_{1.0}$  – условные индексы влажности почвы на глубинах 0,1 м, 0,5 м и 1 м;

$T_{air}$ , – температура воздуха;

$T_{soil}$ , — температура почвы;

$pH$  – показатель кислотности почвы;

$flag$  – признак корректности записи.

После преобразования сигналов датчиков и присвоения служебных параметров устройство сбора и передачи данных формирует строку табличного файла. Такая структура удобна для последующей проверки, агрегации и импорта данных в программную среду моделирования. Пример структуры записи, сохраняемой в файле формата .xls, приведен в таблице 3.2.

Таблица 3.2 – Структура записи данных с устройства сбора и передачи данных

Дата	Время	Номер поля	Влажн ость 0,1 м, усл. ед.	Влажно сть 0,5 м, усл. ед.	Влажн ость 1 м, усл. ед.	Темпе ратур а возду х, °С	Темпер атура почвы, °С	рН почвы	Статус записи
26.05.20 24	09.00	Поле 1	2,00	2,45	2,80	18	16,9	6,8	коррект но

Значения влажности почвы в таблице представлены в условных единицах. Они получены не непосредственно с датчика в готовом виде, а после преобразования аналогового сигнала в цифровой код АЦП и последующего калибровочного пересчета. Такой формат позволяет представить данные с разных глубин в единой табличной структуре и использовать их при формировании локальной базы признаков.

Показатель рН почвы включе

н в структуру записи как дополнительный локальный агрохимический параметр, характеризующий состояние почвенной среды конкретного участка. В отличие от температуры воздуха, для которой была выполнена сопоставительная оценка с данными Казгидромета, верификация рН-канала требует отдельного лабораторного сопоставления с результатами агрохимического анализа почвы. Поэтому в рамках настоящего раздела рН-данные использовались после проверки допустимого диапазона значений и включались в общую таблицу локальных признаков как дополнительный показатель.

Поскольку устройство выполняло измерения каждые 3 часа, исходная таблица имела высокую временную детализацию. Однако урожайность определяется по итогам вегетационного периода, поэтому трехчасовые измерения не рассматривались как самостоятельные строки обучающей выборки. Для согласования данных, полученных с устройства сбора и передачи данных, с сезонным значением урожайности была выполнена агрегация первичных измерений до суточных, месячных и сезонных признаков. Такой подход соответствует практике полевого прогнозирования урожайности, где динамические признаки температуры, влажности и погодных условий предварительно обобщаются по временным интервалам для дальнейшего использования в моделях машинного обучения [10].

Среднесуточное значение параметра рассчитывалось по формуле:

$$\bar{x}_d = \frac{1}{m} \sum_{i=1}^m x_{d,i}, \quad (3.10)$$

где  $\bar{x}_d$  – среднесуточное значение параметра за сутки d;

$x_{d,i}$  – отдельное измерение в течение суток;

m – количество измерений в сутки.

Поскольку устройство выполняло измерения с интервалом 3 часа, количество измерений в сутки составляло:  $m=8$ .

$$\bar{x}_M = \frac{1}{D} \sum_{d=1}^D \bar{x}_d, \quad (3.11)$$

где  $\bar{x}_M$  – среднемесячное значение параметра;  
 $D$  – количество суток в рассматриваемом месяце;  
 $\bar{x}_d$  – среднесуточное значение параметра.

Дополнительно для каждого измеряемого параметра были рассчитаны минимальные и максимальные значения, а также амплитуда изменения за рассматриваемый период:

$$A = x_{max} - x_{min}, \quad (3.12)$$

где  $A$  – амплитуда изменения параметра;  
 $x_{max}$  – максимальное значение параметра за рассматриваемый период;  
 $x_{min}$  – минимальное значение параметра за рассматриваемый период.

После расчета суточных, месячных и сезонных характеристик исходные трехчасовые измерения были приведены к формату агрегированных признаков. В отличие от исходной строки измерения, такая запись отражает не отдельный момент наблюдения, а обобщенное состояние поля за выбранный временной интервал. В состав агрегированных признаков включались средние значения влажности почвы по глубинам, средняя температура воздуха и почвы, минимальные и максимальные значения, а также амплитуда изменения параметров. Именно такие показатели далее использовались при подготовке входного вектора локальной нейросетевой модели.

В результате обработки была сформирована локальная таблица признаков, в которой строки соответствуют объектам анализа, а столбцы содержат агрегированные показатели, рассчитанные на основе данных, полученных с устройства сбора и передачи данных. Данная таблица была объединена с архивными производственными данными ТОО «Агрофирма «Мамлютская», включая фактическую урожайность, и использована на следующем этапе исследования для построения локальной нейросетевой модели прогнозирования урожайности. Предложенный алгоритм обеспечил переход от первичных сигналов датчиков к структурированной базе данных, пригодной для интеллектуального анализа.

### **Выводы по разделу**

В третьей главе показано, что переход от регионального прогнозирования урожайности к локальному сбору полевых данных является необходимым условием для повышения детализации входной информации. Региональная нейросетевая модель позволяет описать общие закономерности формирования урожайности пшеницы в условиях Северо-Казахстанской области, однако для прогноза на уровне

конкретного хозяйства требуется учет фактических почвенно-микrokлиматических условий отдельного посевного участка.

Установлено, что предложенная структура устройства сбора и передачи данных обеспечивает полный цикл работы с локальными полевыми измерениями: регистрацию параметров почвенно-микrokлиматической среды, первичную обработку сигналов, присвоение временной метки, сохранение данных и их передачу на стационарный модуль. Наличие переносного автономного модуля и стационарного модуля приема данных позволяет организовать сбор информации непосредственно в полевых условиях и подготовить ее для дальнейшего анализа.

Подтверждена реализуемость аппаратного решения в виде опытного образца устройства, пригодного для эксплуатации на открытом посевном участке. Использование микроконтроллера, датчиков влажности почвы, температуры воздуха и почвы, рН-канала, модуля памяти, GSM/GPRS-связи и автономной системы питания обеспечивает регистрацию и передачу локальных данных без постоянного подключения к электрической сети. Конструктивное исполнение устройства позволяет размещать его непосредственно на поле в течение периода наблюдений.

Полевая апробация подтвердила пригодность устройства для регистрации локальной динамики температуры воздуха. Сопоставление данных устройства с данными Казгидромета показало высокую согласованность временных рядов: коэффициент корреляции Пирсона составил  $r = 0,97$ , среднеквадратическая ошибка –  $RMSE = 1,3$  °C, средняя абсолютная ошибка –  $MAE = 1,0$  °C. Полученные значения подтверждают возможность использования данных устройства при формировании локальных признаков для нейросетевой модели прогнозирования урожайности. Показано, что измерение влажности почвы на глубинах 0,1 м, 0,5 м и 1 м позволяет выявить различия в динамике водного режима по почвенным горизонтам. Верхний слой характеризуется наибольшей изменчивостью, средний слой изменяется более плавно, а глубокий горизонт сохраняет наибольшую устойчивость. Это подтверждает практическую значимость послойного измерения влажности почвы для оценки условий формирования урожайности на конкретном поле.

Установлено, что данные, получаемые с устройства сбора и передачи данных, требуют предварительного преобразования и структурирования перед использованием в нейросетевой модели. Аналоговые сигналы датчиков влажности почвы преобразуются встроенным АЦП микроконтроллера в цифровой код, после чего пересчитываются по калибровочной зависимости в условные единицы влажности почвы. Каждому измерению присваиваются дата, время, номер поля и глубина установки датчика, что обеспечивает формирование структурированной строки табличного файла.

Показано, что трехчасовые измерения не могут рассматриваться как самостоятельные строки обучающей выборки, поскольку целевая переменная – урожайность – определяется по итогам вегетационного периода. Для согласования локальных измерений с сезонным значением урожайности обеспечена агрегация данных до суточных, месячных и сезонных признаков. В результате сформирована информационная основа для последующего построения локальной нейросетевой модели и адаптации региональной модели к условиям конкретного хозяйства.

## **4 ИНТЕГРАЦИЯ ПОЛЕВЫХ ДАННЫХ В НЕЙРОСЕТЕВУЮ СИСТЕМУ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ**

### **4.1 Формирование, предварительная обработка и разведочный анализ локальных полевых данных**

Современное цифровое земледелие рассматривает данные как основу принятия агротехнологических решений. В концепции Smart Farming данные используются не только для наблюдения за состоянием поля, но и для построения управляемого производственного контура, включающего сбор, хранение, обработку, анализ и применение результатов при принятии решений [66]. В работах, посвященных Agriculture 5.0, подчеркивается, что управление сельскохозяйственным производством строится как последовательность этапов: получение данных в поле, их передача, обработка, анализ и использование для оптимизации агротехнологических операций [67].

В рамках настоящего исследования данная логика используется для перехода от регионального нейросетевого прогнозирования к локальной адаптации модели. Региональная модель, рассмотренная во второй главе, позволяет выявить общие закономерности формирования урожайности пшеницы по 13 районам Северо-Казахстанской области. Однако для конкретного хозяйства такие данные имеют ограничение, поскольку они не всегда отражают состояние отдельного поля, особенности агротехнологий, локальный режим влагообеспеченности и микроклиматические условия. Поэтому в четвертой главе основное внимание уделяется формированию локальной базы данных, пригодной для построения локальной нейросетевой модели.

В систематических обзорах по применению машинного обучения в сельском хозяйстве отмечается, что задачи прогнозирования урожайности обычно решаются на основе совокупности погодных, почвенных, агрохимических и биологических факторов, а искусственные нейронные сети относятся к числу распространенных методов моделирования таких зависимостей [68-70]. Для прогнозирования урожайности пшеницы также показана результативность объединения данных различных источников, включая почвенные измерения, сенсорные данные и характеристики состояния культуры [71]. Это подтверждает необходимость не только построения модели, но и предварительной систематизации входных признаков.

В настоящем исследовании локальная база данных формировалась с использованием исторических данных хозяйства, агрономических отчетов, агрохимических сведений, а также данных, получаемых с устройства сбора и передачи агрометеорологической информации. В третьей главе было показано, что устройство обеспечивает получение локальных полевых данных, которые могут быть преобразованы, проверены, агрегированы и использованы как основа для локальной нейросетевой модели. При этом данные с датчиков влажности почвы не интерпретируются как прямое измерение запаса продуктивной влаги. Датчик формирует электрический сигнал, связанный с влажностным состоянием почвы, а после обработки этот сигнал используется как локальный сенсорный показатель влажности почвы в условных единицах.

Общая схема интеграции локальных полевых данных в нейросетевую систему прогнозирования урожайности представлена на рисунке 4.1.

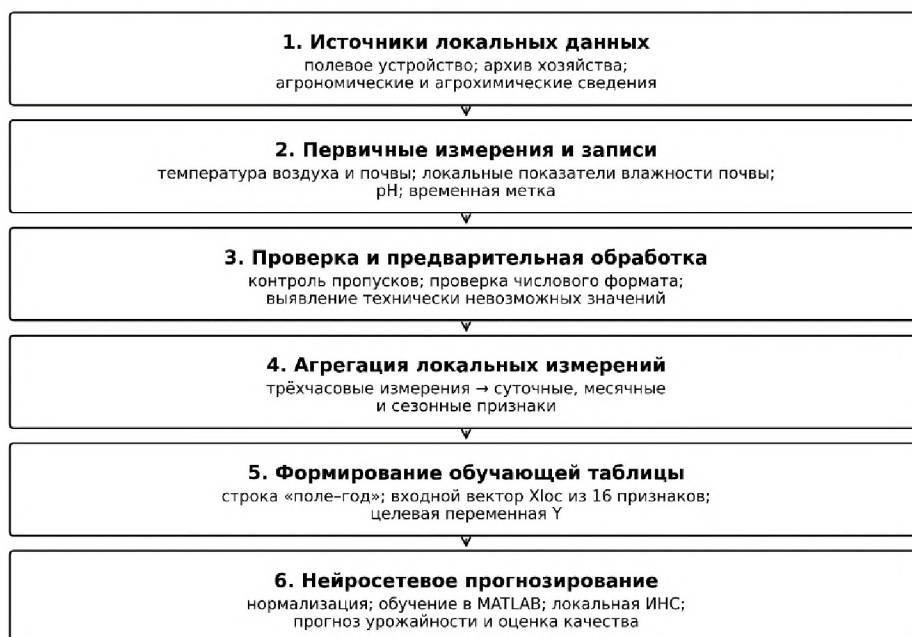


Рисунок 4.1 – Общая схема интеграции локальных полевых данных в нейросетевую систему прогнозирования урожайности

На рисунке 4.1 отражен информационный путь данных от первичного получения локальных измерений до их использования в нейросетевой модели. В отличие от аппаратной схемы устройства, данная схема показывает не конструктивные элементы, а последовательность преобразования данных: получение локальных измерений, их проверку, агрегацию, формирование обучающей таблицы, нормализацию, обучение модели и получение прогноза урожайности. Фитосанитарные показатели локальной базы данных формировались на основе сведений Северо-Казахстанской областной карантинной лаборатории Комитета государственной инспекции в агропромышленном комплексе Министерства сельского хозяйства Республики Казахстан. К данной группе признаков относятся сведения о распространении вредителей, болезней и сорной растительности, используемые в модели как фитосанитарные входные параметры. Агрохимические признаки локальной модели формировались по результатам лабораторного анализа почвенных образцов. В их состав включены показатели содержания азота, фосфора и кислотности почвы, характеризующие агрохимическое состояние исследуемого участка.

В результате объединения данных полевого устройства, исторических сведений хозяйства, фитосанитарной информации и результатов агрохимического анализа формируется локальная таблица наблюдений. Каждая строка такой таблицы соответствует отдельному наблюдению типа «поле-год»

или «участок–год»), а столбцы содержат совокупность входных признаков и целевую переменную – фактическую урожайность пшеницы.

Таблица 4.1 – Источники локальных данных и их роль в нейросетевой модели

Группа данных	Параметры	Источник данных	Форма представления	Использование в модели
Агрометеорологические признаки	Температура воздуха, осадки	Полевое устройство / архивные данные хозяйства	°С, мм	Входные признаки, характеризующие погодные условия вегетационного периода
Почвенные локальные признаки	Локальные показатели влажности почвы на глубинах 0,1 м; 0,5 м; 1 м	Устройство сбора и передачи данных	Условные единицы после калибровочного пересчета	Входные признаки, отражающие влажностное состояние почвы
Дополнительные почвенные признаки	Температура почвы, рН почвы	Полевое устройство / лабораторные данные	°С, рН	Дополнительные признаки состояния почвенной среды
Агрохимические признаки	Содержание азота, содержание фосфора, кислотность почвы	Результаты лабораторного анализа почвенных образцов	Принятая форма лабораторного представления	Входные признаки, характеризующие агрохимическое состояние участка
Фитосанитарные признаки	Сведения о вредителях, болезнях и сорности посевов	Северо-Казахстанская областная карантинная лаборатория	Индексная / балльная оценка	Фитосанитарные входные признаки
Целевая переменная	Урожайность пшеницы	Исторические данные хозяйства	ц/га	Выходной параметр нейросетевой модели

Представленная структура локальной базы данных обеспечивает объединение разнородных источников информации в единую обучающую таблицу. Входные признаки характеризуют агрометеорологические, почвенные, агрохимические и фитосанитарные условия формирования урожайности, а целевая переменная представлена фактической урожайностью пшеницы в ц/га. В таком виде локальная база данных может быть использована для обучения, проверки и последующей оценки локальной нейросетевой модели.

Перед построением локальной нейросетевой модели была выполнена проверка качества сформированной базы данных. На данном этапе контролировалось наличие значений по входным признакам и целевой переменной, корректность числового формата, а также логическая допустимость диапазонов измеряемых параметров. Особое внимание уделялось целевой переменной – урожайности пшеницы, поскольку наблюдения с отсутствующим значением урожайности не могут использоваться при обучении модели. Значения, имеющие признаки технических ошибок, сбоя датчика, нарушения записи или передачи данных, подлежали дополнительной проверке и исключались из дальнейшего анализа либо корректировались на основании допустимых правил обработки данных.

Перед построением локальной нейросетевой модели была выполнена проверка качества сформированной базы данных. На данном этапе контролировалось наличие значений по входным признакам и целевой переменной, корректность числового формата, а также логическая допустимость диапазонов измеряемых параметров. Особое внимание уделялось целевой переменной – урожайности пшеницы, поскольку наблюдения с отсутствующим значением урожайности не могут использоваться при обучении модели. Значения, имеющие признаки технических ошибок, сбоя датчика, нарушения записи или передачи данных, подлежали дополнительной проверке и исключались из дальнейшего анализа либо корректировались на основании допустимых правил обработки данных.

Отдельное внимание уделялось согласованию временного масштаба локальных измерений с целевой переменной. Первичные измерения устройства фиксируются с заданным временным интервалом, в том числе с интервалом 3 часа, тогда как урожайность является итоговым сезонным показателем. Поэтому отдельные трехчасовые измерения не использовались как самостоятельные строки обучающей выборки, а агрегировались до суточных, месячных и сезонных признаков. Такой подход соответствует логике формирования признаков в задачах прогнозирования урожайности, где входная информация должна быть приведена к масштабу, сопоставимому с целевой переменной [5, 7].

В общем виде суточное среднее значение локального признака может быть записано следующим образом:

$$\bar{x}_d = \frac{1}{m} \sum_{j=1}^m x_{dj}, \quad (4.1)$$

где  $\bar{x}_d$  – среднее значение признака за сутки,

$x_{dj}$  – j-е измерение в течение суток,

(m) – число измерений за сутки.

При интервале измерений 3 часа (m = 8).

Далее суточные значения могут быть агрегированы до месячного или сезонного уровня:

$$\bar{x}_{month} = \frac{1}{n} \sum_{d=1}^n \bar{x}_d, \quad (4.2)$$

где  $\bar{x}_{month}$  – среднее значение признака за месяц,  
 (n) – число суток в месяце, по которым имеются корректные измерения.

Схема формирования локальной обучающей выборки представлена на рисунке 4.2.



Рисунок 4.2 – Схема формирования локальной обучающей выборки

Рисунок 4.2 отражает переход от первичных трехчасовых измерений к агрегированным признакам, сопоставимым с сезонной урожайностью. В результате формируется строка наблюдения типа «поле-год» или «участок-год», которая затем используется как входной вектор локальной нейросетевой модели.

$$X_{loc} = \{x_1, x_2, \dots, x_{16}\}, \quad (4.3)$$

где  $x_1, x_2, \dots, x_{16}$  – агрегированные и нормализованные локальные признаки, характеризующие агрометеорологические, почвенные, агрохимические и фитосанитарные условия формирования урожайности.

Для оценки пригодности локальной базы к нейросетевому моделированию выполнялся разведочный анализ данных. В отличие от второй главы, где EDA применялся к региональной базе, в данном подразделе разведочный анализ имеет

более узкую цель – проверить качество локальной выборки и определить, может ли она быть использована как обучающий массив для локальной модели.

Разведочный анализ локальной базы включал проверку структуры таблицы, оценку полноты данных, анализ диапазонов значений, расчет описательной статистики и анализ распределения урожайности. Такой подход позволяет перейти от исходной разнородной информации к структурированному массиву, пригодному для машинного обучения [72].

Описательная статистика локальных признаков представлена в таблице 4.2.

Таблица 4.4 – Описательная статистика локальных признаков

Переменная	Признак	Ед. изм.	Среднее значение	Стандартное отклонение	CV, %	Min	Max
X <sub>1</sub>	Локальный показатель влажности почвы 0,1 м	усл. ед.	11,47	4,27	37,3	2,4	27,6
X <sub>2</sub>	Локальный показатель влажности почвы 0,5 м	усл. ед.	58,54	16,16	27,6	22,6	109,6
X <sub>3</sub>	Локальный показатель влажности почвы 1 м	усл. ед.	114,58	28,74	25,1	47,3	191,3
X <sub>4</sub>	Содержание азота в почве	мг/кг	49,61	6,71	13,5	31,9	92,7
X <sub>5</sub>	Содержание фосфора в почве	мг/кг	18,79	5,64	30,0	6,1	53,3
X <sub>6</sub>	Осадки за май	мм	40,51	22,01	54,3	4,0	100,9
X <sub>7</sub>	Осадки за июнь	мм	35,11	26,56	75,6	3,3	156,4
X <sub>8</sub>	Осадки за июль	мм	72,11	42,36	58,7	5,1	285,0
X <sub>9</sub>	Осадки за август	мм	45,90	27,85	60,7	4,9	145,2
X <sub>10</sub>	Температура воздуха за май	°С	13,19	1,24	9,4	6,9	15,7
X <sub>11</sub>	Температура воздуха за июнь	°С	18,69	1,52	8,1	12,5	23,2

X <sub>12</sub>	Температура воздуха за июль	°С	19,17	1,94	10,1	11,2	24,0
X <sub>13</sub>	Температура воздуха за август	°С	17,76	1,66	9,3	14,0	21,8
X <sub>14</sub>	Индекс болезней	балл	2,47	0,87	35,3	1,0	4,0
X <sub>15</sub>	Индекс сорности	балл	1,92	0,18	9,5	1,5	3,0
X <sub>16</sub>	Индекс вредителей	балл	1,55	0,57	36,9	0,0	2,6
Y	Урожайность пшеницы	ц/га	14,34	3,70	25,8	5,9	25,5

Анализ таблицы 4.2 показывает, что локальная база содержит признаки с различными единицами измерения и существенно отличающимися диапазонами значений. Наибольшая изменчивость наблюдается по осадкам за июнь, июль и август, что отражает высокую межгодовую вариабельность влагообеспечения в период вегетации. Значительная вариация также характерна для фитосанитарных признаков, особенно индекса вредителей и индекса болезней, что указывает на неодинаковую выраженность биотических факторов по годам и участкам наблюдений. Урожайность пшеницы изменяется от 5,9 до 25,5 ц/га при среднем значении 14,34 ц/га, что отражает межгодовую и локальную вариабельность условий формирования урожая.

Таблица описательной статистики подтверждает необходимость нормализации данных перед обучением нейросетевой модели. Это связано с тем, что входные признаки имеют разные единицы измерения и различные числовые диапазоны: температура выражается в градусах, осадки – в миллиметрах, агрохимические показатели – в лабораторных единицах представления, а фитосанитарные признаки – в балльной или индексной шкале. Нормализация применяется не для изменения физического смысла данных, а для приведения признаков к сопоставимому числовому масштабу.

Дополнительно для локальной базы было проанализировано распределение целевой переменной – урожайности пшеницы. Для сохранения единого стиля представления результатов графический анализ выполнен по той же логике, что и во второй главе: построена гистограмма распределения урожайности, а также диаграмма размаха по типам лет.

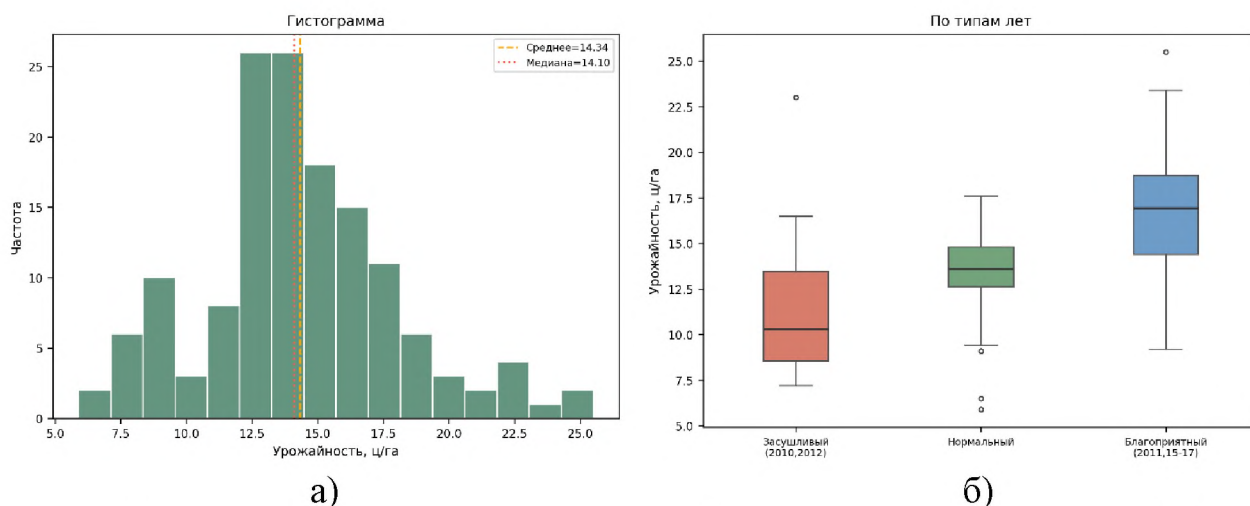


Рисунок 4.3 – Распределение урожайности пшеницы в локальной базе данных:  
 а) гистограмма;  
 б) распределение урожайности по типам лет

На рисунке 4.3а представлена гистограмма распределения фактической урожайности пшеницы по 143 наблюдениям локальной базы данных. Среднее значение урожайности составляет 14,34 ц/га, медиана – 14,10 ц/га, стандартное отклонение — 3,70 ц/га. Близость среднего значения и медианы свидетельствует об отсутствии резко выраженной асимметрии распределения. Значения урожайности изменяются от 5,90 до 25,50 ц/га, что отражает межгодовую и локальную вариабельность условий формирования урожая.

На рисунке 4.3б урожайность дополнительно представлена по типам лет. В засушливые годы среднее значение урожайности составило 11,35 ц/га, в нормальные годы – 13,42 ц/га, в благоприятные годы – 16,98 ц/га. Такая дифференциация подтверждает, что локальная база данных отражает различия агроклиматических условий и содержит наблюдения как для неблагоприятных, так и для более продуктивных лет. Это важно для последующего обучения нейросетевой модели, поскольку модель должна учитывать не только средние условия, но и вариативность урожайности при различных сценариях вегетационного периода.

Полученный характер распределения позволяет использовать урожайность в исходной шкале измерения без дополнительного логарифмического преобразования целевой переменной. При этом наличие различий между типами лет подтверждает необходимость включения в модель агрометеорологических, почвенных, агрохимических и фитосанитарных признаков, отражающих условия формирования урожая.

Корреляционный анализ локальных признаков использовался как вспомогательный инструмент предварительной диагностики данных. Поскольку общий подход к построению и интерпретации корреляционной матрицы был подробно рассмотрен во второй главе, в настоящем разделе основное внимание уделено описательной статистике, распределению урожайности и подготовке локальной базы к обучению нейросетевой модели. При этом следует учитывать,

что низкая парная корреляция отдельного признака с урожайностью не является основанием для его исключения, поскольку нейросетевая модель способна выявлять нелинейные взаимодействия между агрометеорологическими, почвенными, агрохимическими и фитосанитарными факторами [68, 73].

По результатам предварительной обработки и разведочного анализа локальная база данных приводится к формату, пригодному для обучения нейронной сети. Строки соответствуют наблюдениям типа «поле–год» или «участок–год», столбцы содержат агрегированные входные признаки, а целевая переменная представлена фактической урожайностью пшеницы в ц/га. Такой формат обеспечивает прямое использование данных в MATLAB при построении локальной нейросетевой модели.

Таким образом, в разделе 4.1 установлено, что локальные полевые данные после проверки, агрегации, нормализации и разведочного анализа могут быть использованы как обучающая база для локальной нейросетевой модели. В отличие от региональной базы, локальная выборка отражает условия конкретного хозяйства, что создает основу для дальнейшей адаптации прогноза урожайности к фактическим условиям поля. Следующий этап исследования связан с разработкой и обучением локальной нейросетевой модели архитектуры 16–10–1.

#### **4.2 Проектирование и обучение локальной нейросетевой модели**

После формирования, проверки и предварительной обработки локальной базы данных следующим этапом исследования является построение локальной нейросетевой модели прогнозирования урожайности пшеницы. В отличие от региональной модели, отражающей обобщенные закономерности по районам Северо-Казахстанской области, локальная модель ориентирована на использование признаков, характеризующих условия конкретного хозяйства или отдельного участка. Это позволяет уточнить прогноз за счет включения данных, полученных непосредственно в полевых условиях.

Задача прогнозирования урожайности в данном разделе рассматривается как задача нелинейной регрессии. Необходимо построить отображение, которое по входному вектору локальных признаков определяет прогнозное значение урожайности пшеницы:

$$Y = f(X_{loc}), \quad (4.4)$$

где  $X_{loc}$  – входной вектор локальных признаков, включающий 16 параметров;

$Y$  – урожайность пшеницы, ц/га;

$f$  – нелинейная функция, аппроксимируемая искусственной нейронной сетью.

Входной вектор модели включает агрометеорологические, почвенные, агрохимические и фитосанитарные признаки, сформированные на предыдущем этапе. В качестве выходного параметра используется фактическая урожайность пшеницы. Таким образом, обучающая выборка имеет вид:

$$D = \{(X_i, Y_i)\}_{i=1}^n, \quad (4.5)$$

где  $X_i$  – вектор входных признаков для  $i$ -го наблюдения;  $Y$   
 $i$  – фактическая урожайность для соответствующего наблюдения;  
 $n=143$  – количество наблюдений в локальной базе данных.

В качестве локальной модели выбрана полносвязная прямонаправленная нейронная сеть типа multilayer perceptron. Данный тип сети широко применяется в задачах регрессионного прогнозирования, поскольку позволяет аппроксимировать нелинейные зависимости между входными признаками и целевой переменной [69-72, 74]. Для рассматриваемой задачи была использована архитектура 16–10–1, включающая 16 входных нейронов, один скрытый слой из 10 нейронов и один выходной нейрон, соответствующий прогнозируемой урожайности.

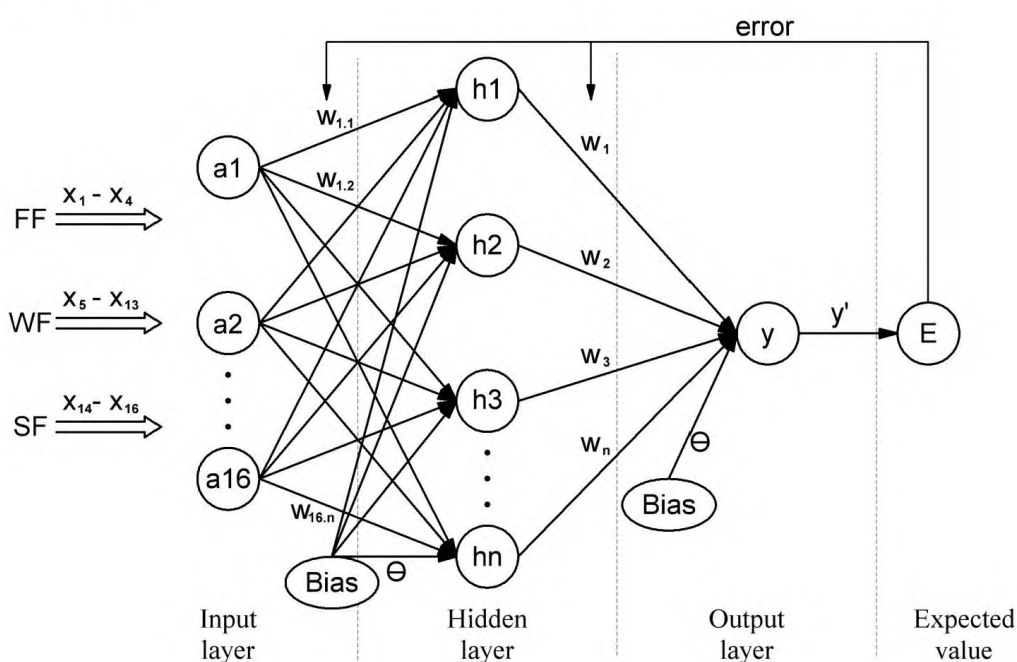


Рисунок 4.4 – Архитектура локальной нейросетевой модели 16–10–1

На рисунке 4.4 должна представлена структура локальной нейросетевой модели: входной слой из 16 признаков, скрытый слой из 10 нейронов и выходной слой, формирующий прогноз урожайности пшеницы. Входные признаки целесообразно сгруппированы по блокам агрометеорологические, почвенные, агрохимические и фитосанитарные параметры.

Для скрытого слоя использовалась функция активации гиперболического тангенса  $\tanh$ , обеспечивающая нелинейное преобразование входных сигналов. На выходном слое применялась линейная функция активации  $\text{relu}$ , что соответствует регрессионной постановке задачи, поскольку урожайность является непрерывной числовой величиной. Такое сочетание функций

активации позволяет сети учитывать нелинейные связи между факторами и при этом формировать прогноз в исходной количественной шкале.

Количество настраиваемых параметров модели 16–10–1 определяется числом весов и смещений между слоями:

$$W = 16 \cdot 10 + 10 + 10 \cdot 1 + 1 = 181, \quad (4.6)$$

где  $16 \cdot 10$  – веса между входным и скрытым слоями;

$10$  – смещения нейронов скрытого слоя;

$10 \cdot 1$  – веса между скрытым и выходным слоями;

$1$  – смещение выходного нейрона. Полученное число параметров является приемлемым для локальной выборки при условии применения нормализации данных и контроля переобучения на валидационной выборке.

Основные параметры локальной нейросетевой модели представлены в таблице 4.3.

Таблица 4.3 – Параметры локальной нейросетевой модели

Параметр	Значение
Тип модели	Полносвязная прямонаправленная нейронная сеть
Назначение модели	Прогнозирование урожайности пшеницы
Количество входных признаков	16
Целевая переменная	Урожайность пшеницы, ц/га
Архитектура	16–10–1
Количество скрытых слоев	1
Количество нейронов скрытого слоя	10
Функция активации скрытого слоя	tansig
Функция активации выходного слоя	purelin
Алгоритм обучения	Levenberg–Marquardt, trainlm
Нормализация данных	mapminmax
Среда моделирования	MATLAB Neural Network Toolbox

До обучения нейронной сети входные признаки и целевая переменная были нормализованы. Необходимость нормализации обусловлена тем, что признаки имеют разные единицы измерения и различные числовые диапазоны. Так, температура выражается в градусах Цельсия, осадки – в миллиметрах, агрохимические показатели – в лабораторных единицах представления, а фитосанитарные признаки – в балльной или индексной шкале. Без масштабирования признаки с большими числовыми значениями могут оказывать непропорциональное влияние на процесс обучения.

Нормализация выполнялась методом min–max с приведением данных к диапазону  $[-1;1]$ , что соответствует стандартному применению функции mapminmax в MATLAB:

$$\bar{x} = 2 \cdot \frac{x - x_{min}}{x_{max} - x_{min}} - 1, \quad (4.7)$$

где  $x$  – исходное значение признака;

$x_{max}$  и  $x_{min}$  – минимальное и максимальное значения признака;

$\bar{x}$  – нормализованное значение признака. Приведение данных к диапазону  $[-1;1]$  согласуется с областью значений функции активации *tansig* и способствует более устойчивому обучению нейронной сети.

Обучение локальной нейросетевой модели выполнялось в среде MATLAB Neural Network Toolbox с использованием алгоритма Левенберга-Марквардта (*trainlm*). Данный алгоритм является модификацией метода наименьших квадратов для нелинейных задач и сочетает свойства метода Гаусса-Ньютона и градиентного спуска [9–11]. В MATLAB функция *trainlm* применяется для обучения нейронных сетей путем обновления весов и смещений в соответствии с оптимизацией Левенберга-Марквардта.

Критерием обучения выступала среднеквадратическая ошибка:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2, \quad (4.8)$$

где  $Y_i$  – фактическое значение урожайности;

$\hat{Y}_i$  – прогнозное значение урожайности;

$n$  – количество наблюдений.

Минимизация MSE позволяет уменьшить среднее квадратическое отклонение прогнозных значений от фактических. Для оценки качества обучения локальная база данных разделялась на обучающую, валидационную и тестовую подвыборки. Обучающая подвыборка использовалась для настройки весовых коэффициентов сети, валидационная – для контроля обобщающей способности и предотвращения переобучения, тестовая — для итоговой независимой оценки точности модели. В рабочем эксперименте использовалось соотношение 70 % / 15 % / 15 %, что соответствует распространенной схеме первичной оценки нейросетевых моделей в MATLAB.

Выбор архитектуры локальной модели выполнялся путем экспериментального сравнения сетей с различным числом нейронов в скрытом слое. Рассматривались архитектуры 16–5–1, 16–7–1, 16–10–1, 16–12–1 и 16–15–1. При сравнении учитывались точность прогноза на тестовой выборке, средняя абсолютная процентная ошибка, число настраиваемых параметров и риск переобучения.

Число параметров для архитектуры с  $h$  нейронами в скрытом слое рассчитывалось по формуле:

$$W = 16h + h + h + 1 = 18h + 1 \quad (4.9)$$

Результаты сравнения архитектур представлены в таблице 4.4.

Таблица 4.4 – Сравнение архитектур локальной нейросетевой модели

Архитектура	Число параметров $W$	$n_{обуч}/W$	$R^2$ на тестовой выборке	MAPE, %	Оценка
16–5–1	91	1,1	0,891	18,5	Недостаточная точность
16–7–1	127	0,79	0,934	15,3	Приемлемо
16–10–1	181	0,55	0,9665	12,02	Оптимальная
16–12–1	217	0,46	0,951	13,8	Начало переобучения
16–15–1	271	0,37	0,927	16,1	Переобучение

Анализ таблицы 4.4 показывает, что увеличение числа нейронов скрытого слоя от 5 до 10 приводит к повышению точности модели. Архитектура 16–5–1 имеет относительно небольшое число параметров, однако не обеспечивает достаточной точности прогноза. Архитектура 16–7–1 демонстрирует улучшение качества, но уступает варианту 16–10–1. Наилучшее сочетание точности и устойчивости показала архитектура 16–10–1, для которой достигнуто значение  $R^2=0,9665$  на тестовой выборке при MAPE = 12,02 %. Дальнейшее увеличение числа нейронов до 12 и 15 не привело к улучшению результата. Напротив, наблюдается снижение  $R^2$  и рост MAPE, что указывает на ухудшение обобщающей способности модели. Это подтверждает, что чрезмерное усложнение архитектуры при ограниченном объеме локальной выборки повышает риск переобучения. Поэтому архитектура 16–10–1 была выбрана в качестве базовой локальной нейросетевой модели.

Таблица 4.5 – Сравнение качества локальных нейросетевых моделей при различном числе нейронов скрытого слоя

Параметр сравнения	Региональная модель	Локальная модель
Архитектура	16–3–1	16–10–1
Алгоритм обучения	trainlm (LM)	trainlm (LM)
Источник метеоданных	МС Мамлютка (одна станция для всех районов)	Полевые датчики CSMSv2.0 (пространственно привязанные)
Период наблюдений	2007–2020	2008–2017
Объем выборки $n$	182 наблюд.	143 наблюд.
Число параметров $W$	55	181
Соотношение $n_{обуч}/W$	$\approx 3,3$	$\approx 0,55$
$R$ (тест)	0,7796	0,9831 (+26,1 п.п.)
$R^2$ (тест)	0,608	0,9665 (+58,8 п.п.)

RMSE, ц/га (тест)	2,890	3,368
MAPE, % (тест)	12,10	12,02 ( $\approx$ идентично)
MSE_val (наилучшее)	7,16 (ц/га) <sup>2</sup>	0,9511 (ц/га) <sup>2</sup> ( $-7,5\times$ )

В таблице 4.5 целесообразно представить изменение ошибки или показателя качества при переходе от 5 к 15 нейронам скрытого слоя. Такой рисунок наглядно покажет, что оптимум достигается при 10 нейронах, а дальнейшее увеличение сложности модели не улучшает результат.

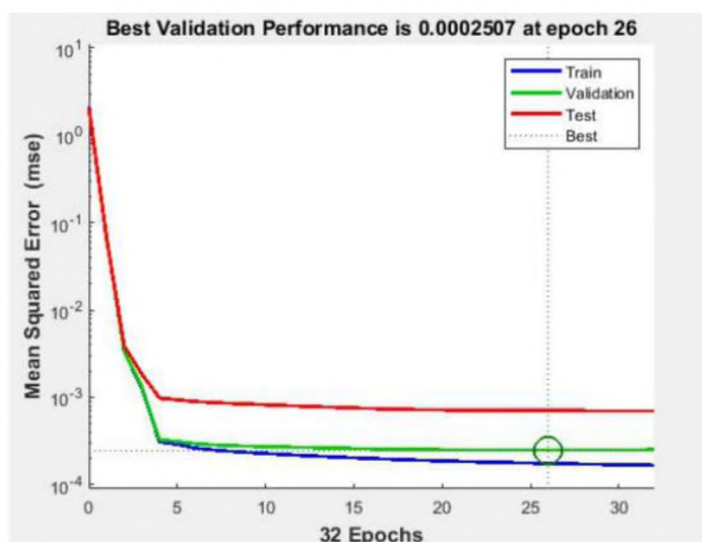


Рисунок 4.6 – Процесс обучения локальной нейросетевой модели в MATLAB

График обучения позволяет оценить динамику изменения ошибки на обучающей, валидационной и тестовой подвыборках. Если валидационная ошибка после некоторого момента перестает уменьшаться или начинает возрастать, обучение останавливается, что предотвращает дальнейшее подстраивание модели под обучающую выборку. В результате сеть сохраняет способность к обобщению на новых данных.

Таким образом, в разделе 4.2 разработана локальная нейросетевая модель прогнозирования урожайности пшеницы на основе 16 входных признаков. В результате сравнения нескольких архитектур установлено, что модель 16–10–1, обученная алгоритмом Левенберга-Марквардта в MATLAB, обеспечивает наилучшее сочетание точности и обобщающей способности. Данная модель принимается в качестве базовой локальной модели для последующей оценки качества прогноза и сравнения с региональной нейросетевой моделью.

### Выводы по разделу

Сформирована локальная выборка данных, включающая 143 наблюдения по 13 районам Северо-Казахстанской области за 2008–2017 гг. В качестве входных параметров использованы 16 предикторов, объединенных в три группы: агрохимические, метеорологические и фитосанитарные. Ключевым источником

данных выступают измерения полевых датчиков разработанного устройства сбора и передачи данных. Переход от интерполированных метеостанционных данных к пространственно привязанным полевым измерениям позволил преодолеть основное ограничение региональной модели и повысить информативность входной выборки.

По результатам систематического перебора пяти архитектур искусственной нейронной сети – от 16–5–1 до 16–15–1– оптимальной признана модель 16–10–1, содержащая 181 настраиваемый параметр. Обучение модели выполнено с использованием алгоритма Левенберга-Марквардта и механизма ранней остановки, что позволило обеспечить устойчивость модели к переобучению при соотношении количества обучающих наблюдений к числу параметров  $n_{\text{обуч}}/W = 0,55$ .

Разработанная локальная нейросетевая модель архитектуры 16–10–1 показала высокую точность прогнозирования урожайности на независимой тестовой выборке:  $R^2 = 0,9665$ , MAPE = 12,02 %, RMSE = 3,368 ц/га. Полученные результаты подтверждены опубликованными научными данными [41]. Высокое значение коэффициента детерминации при сопоставимом уровне MAPE свидетельствует о низком систематическом разбросе прогнозных значений и высокой информативности локальных полевых данных.

Сравнительный анализ региональной и локальной моделей показал, что переход от региональных агрометеорологических данных к локальным полевым измерениям обеспечивает повышение коэффициента детерминации с 0,608 до 0,9665, то есть на 58,8 процентных пункта, при одновременном снижении валидационной ошибки MSE\_val в 7,5 раза. Данный результат количественно подтверждает центральную гипотезу исследования о решающей роли пространственно привязанного мониторинга в задаче нейросетевого прогнозирования урожайности зерновых культур.

На основе полученных результатов разработана трехэтапная методология адаптации нейросетевой модели с применением принципов трансферного обучения. На первом этапе используется региональная модель с гибридным входным вектором, обеспечивающая прогноз с MAPE на уровне 20–30 %. На втором этапе вводится поправочный коэффициент  $k$ , позволяющий снизить ошибку прогноза до 15–20 %. На третьем этапе предусматривается локальная адаптация модели за счет fine-tuning или взвешенного переобучения, что позволяет достичь MAPE на уровне 10–15 %. Предложенная методология может применяться уже с первого вегетационного сезона эксплуатации аппаратного комплекса и обеспечивает поэтапное автоматическое повышение точности прогноза по мере накопления локальных данных.

## ЗАКЛЮЧЕНИЕ

В диссертационной работе согласно поставленным цели и задач, была разработана нейросетевая система прогноза урожайности зерновых культур.

Основные научные и практические результаты диссертационной работы заключаются в следующем.

Экспериментальным путем определен набор входных данных искусственной нейронной сети для прогнозирования урожайности зерновых культур в СКО.

В ходе работы была разработана система прогнозирования урожайности яровой пшеницы для Северо-Казахстанской области на основе искусственной нейронной сети. Эта система учитывает такие важные факторы, как количество продуктивной влаги в почве, погодные условия, плодородие почвы, а также наличие вредителей, болезней и сорняков. Благодаря гибкости и адаптивности модель может быть использована для прогнозирования урожайности других сельскохозяйственных культур и в других регионах.

Выявлены и проанализированы основные агроклиматические, агрохимические и фитосанитарные факторы, оказывающие влияние на формирование урожайности пшеницы.

Разработаны методы мониторинга значимых факторов, заключающиеся в определении наиболее оптимальных способов, условий и интервалов измерения для получения наиболее точного прогноза.

Создана комплексная модель прогнозирования урожайности и валового сбора на основе искусственной нейронной сети, которая показала высокую точность в прогнозировании урожайности на основе исторических данных.

Проведена оценка точности предложенной модели с использованием исторических данных агроклиматических и агрохимических факторов, а также данных по урожайности конкретного поля, что позволило убедиться в ее достоверности и применимости на практике.

Разработано автономное устройство собирающее и передающее агрометеорологические данные с исследуемого поля для дальнейшего построения прогнозной модели. Система состоит из 4 модулей, устанавливаемых непосредственно на исследуемом поле и одного модуля, производящего прием собираемых данных по беспроводному радиоканалу. Собираемые данные адаптированы под разработанную нейронную сеть. Система не требует обслуживания имеет полное автономное питание и устанавливается вначале посева и производит сбор данных в течении всего вегетационного периода.

Результаты данного исследования подтвердили эффективность применения искусственных нейронных сетей для решения задач прогнозирования в сельском хозяйстве. Разработанная система может стать полезным инструментом для фермеров и агропромышленных компаний, позволяя улучшить управление урожайностью и планирование агротехнических мероприятий.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Phalan, B., Green, R., & Balmford, A. (2014). Closing yield gaps: perils and possibilities for biodiversity conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1639), 20120285. <https://doi.org/10.1098/rstb.2012.0285>
2. Tilman, D., Balzer, C., Hill, J., & Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the national academy of sciences*, 108(50), 20260-20264. <https://doi.org/10.1073/pnas.1116437108>
3. Basso, B., Liu, L., 2019. Seasonal crop yield forecast: methods, applications, and accuracies. *Advances in Agronomy* 154, 201–255.
4. Ben-Ari, T., et al., 2018. Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nat. Commun.* 9 (1), 1627. <https://www.nature.com/articles/s41467-018-04087-x>
5. Funk, C., et al., 2019b. Recognizing the famine early warning systems network: over 30 years of drought early warning science advances and partnerships promoting global food security. *Bull. Am. Meteorol. Soc.* 100 (6), 1011–1027. <http://doi.org/10.1175/bams-d-17-0233.1>
6. Headey, D., 2011. Rethinking the global food crisis: the role of trade shocks. *Food Policy* 36 (2), 136–146. <http://dx.doi.org/10.1016/j.foodpol.2010.10.003>
7. Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sens. Environ.* 141, 116–128. <http://dx.doi.org/10.1016/j.rse.2013.10.027>
8. MacDonald, R.B., Hall, F.G., 1980. Global crop forecasting. *Science* 208, 670–679. <http://dx.doi.org/10.1126/science.208.4445.670>
9. Puma, M.J., Bose, S., Chon, S.Y., Cook, B.I., 2015. Assessing the evolving fragility of the global food system. *Environ. Res. Lett.* 10 (2), 024007. <http://dx.doi.org/10.1088/1748-9326/10/2/024007>
10. Stone, R.C., Meinke, H., 2005. Operational seasonal forecasting of crop performance. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 360 (1463), 2109–2124. <https://doi.org/10.1098/rstb.2005.1753>
11. Fischer, R., 2015. Definitions and determination of crop yield, yield gaps, and of rates of change. *Field Crop Res.* 182, 9–18. <http://dx.doi.org/10.1016/j.fcr.2014.12.006>
12. Nandram, B., Berg, E., Barboza, W., 2014. A hierarchical Bayesian model for forecasting state-level corn yield. *Environ. Ecol. Stat.* 21, 507–530. <http://dx.doi.org/10.1007/s10651-013-0266-z>
13. Pease, J.W., Wade, E.W., Skees, J.S., Shrestha, C.M., 1993. Comparisons between subjective and statistical forecasts of crop yields. *Rev. Agric. Econ.* 15, 339–350. <https://doi.org/10.2307/1349453>
14. Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sens. Environ.* 141, 116–128. <http://dx.doi.org/10.1016/j.rse.2013.10.027>

15. Lobell, D.B., Schlenker, W., Costa-Roberts, J., 2011. Climate trends and global crop production since 1980. *Science* 333, 616–620. <https://doi.org/10.1126/science.1204531>
16. Arkin, G.F., Maas, S.J., Richardson, C.W., 1980. Forecasting grain Sorghum yields using simulated weather data and updating techniques. *Trans. ASAE* 23, 676–680.
17. Kadaja, J., Saue, T., Vii, P., 2009. Probabilistic yield forecast based on a production process model. In: Li, D., Zhao, C. (Eds.), *Computer and Computing Technologies in Agriculture II*, Vol. 1. vol. 293. Springer, pp. 487–494. [http://doi.org/10.1007/978-1-4419-0209-2\\_50](http://doi.org/10.1007/978-1-4419-0209-2_50)
18. Reynolds, C.A., Yitayew, M., Slack, D.C., Hutchinson, C.F., Huete, A., Petersen, M.S., 2000. Estimating crop yields and production by integrating the FAO crop specific water balance model with real-time satellite data and ground-based ancillary data. *Int. J. Remote Sens.* 21, 3487–3508. <https://doi.org/10.1080/014311600750037516>
19. Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Comput. Electron. Agric.* 151, 61–69. <https://doi.org/10.1016/j.compag.2018.05.012>
20. Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press. <https://www.deeplearningbook.org/> (Last accessed: May 11, 2022).
21. Tian, Hui ren, et al. "An IPSO-BP neural network for estimating wheat yield using two remotely sensed variables in the Guanzhong Plain, PR China." *Computers and Electronics in Agriculture* 169 (2020): 105180. <https://doi.org/10.1016/j.compag.2019.105180>
22. Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P., Lichtenberger, J., 2018. Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. *Agric. Meteorol.* 260, 300–320. <http://dx.doi.org/10.1016/j.agrformet.2018.06.009>
23. Singh, P.K., Singh, K.K., Singh, P., Balasubramanian, R., Baxla, A.K., Kumar, B., Gupta, A., Rathore, L.S., Kalra, N., 2017. Forecasting of wheat yield in various agro-climatic regions of Bihar by using CERES-wheat model. *J. Agrometeorol.* 19, 346–349. <https://doi.org/10.54386/jam.v19i4.604>
24. Portmann, F.T., Siebert, S., Doll, P., 2010. MIRCA2000—global monthly irrigated and drained crop areas around the year 2000: a new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochem. Cycles* 24GB1011. <https://doi.org/10.1029/2008GB003435>
25. Giri, A.K., Bhan, M., Agrawal, K.K., 2017. Districtwise wheat and Rice yield predictions using meteorological variables in eastern Madhya Pradesh. *J. Agrometeorol.* 19, 366–368. <https://doi.org/10.54386/jam.v19i4.610>
26. Singh, M., Sharma, S., 2017. Forecasting the maize yield in Himachal Pradesh using climatic variables. *J. Agrometeorol.* 19, 167–169. <https://doi.org/10.54386/jam.v19i2.715>

27. Schauburger, B., Gornott, C., Wechsung, F., 2017. Global evaluation of a semi-empirical model for yield anomalies and application to within-season yield forecasting. *Glob. Chang. Biol.* 23, 4750–4764. <https://doi.org/10.1111/gcb.13738>
28. Caselli, M.; Trizio, L.; de Gennaro, G.; Ielpo, P. A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model. *Water. Air. Soil Pollut.* **2009**, 201, 365–377.], Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agric. For. Meteorol.* 274, 144–159. <http://dx.doi.org/10.1007/s11270-008-9950-2>
29. Niedbała, G.; Kurasiak-Popowska, D.; Stuper-Szablewska, K.; Nawracała, J. Application of artificial neural networks to analyze the concentration of ferulic acid, deoxynivalenol, and nivalenol in winter wheat grain. *Agriculture* **2020**, 10, 127 <http://dx.doi.org/10.3390/agriculture10040127>
30. Zaefizadeh, M.; Jalili, A.; Khayatnezhad, M.; Gholamin, R.; Mokhtari, T. Comparison of multiple linear regressions (MLR) and artificial neural network (ANN) in predicting the yield using its components in the hulless barley. *Adv. Environ. Biol.* **2011**, 5, 109–113 <https://doi.org/10.1081/ddc-120018209>
31. Агентство Республики Казахстан по статистике, статистический сборник. Портрет села. – Астана, 2011. – 92 с. 33. <https://stat.gov.kz/api/getFile/?docId=WC16200032726>
32. Грибский А. А. Почвы и земельные ресурсы Северо-Казахстанской области. – Петропавловск, 2005. – 123 с.] 34.
33. Ritchie, S. W., J. J. Hanway, and H. E. Thompson. 1996. How a soybean plant develops. Special Report 53. Ames, Iowa: Iowa State University Cooperative Extension Service. <https://dr.lib.iastate.edu/entities/publication/b8b1c46c-e131-4718-bd9921c3071c1f0a>
34. Растениеводство (К.К.Аринов, К.М.Мусынов, Н.А. Шестакова, А.А.Серекпаев. Астана, Типография АО «КазАТУ им. С.Сейфуллина», 2013,507с
35. Сайт Бюро национальной статистики Агентства по стратегическому планированию и реформам Республики Казахстан <https://www.gov.kz/memleket/entities/stat?lang=ru> (Last accessed: December 16, 2021)
36. MathWorks. 2005. Neural network toolbox. Natick, Mass. The MathWorks, Inc. Available at: [www.mathworks.com/products/neuralnet/](http://www.mathworks.com/products/neuralnet/)
37. Drummond, S. T., K. A. Sudduth, and S. J. Birrell. 1995. Analysis and correlation methods for spatial data. ASAE Paper No. 951335. St. Joseph, Mich.: ASAE. <http://dx.doi.org/10.13031/2013.20464>
38. Irmak, A., et al. "Analysis of spatial yield variability using a combined crop model-empirical approach." *Transactions of the ASABE* 49.3 (2006): 811-818. <http://dx.doi.org/10.13031/2013.20464>
39. Wilkerson, J. B., R. Sui, W. E. Hart, L. R. Wilhelm, and D. D. Howard. 1999. Artificial neural networks for determining nitrogen status in corn. ASAE Paper No. 993042. St. Joseph, Mich.: ASAE

40. Braga, R. P. 2000. Predicting the spatial pattern of grain yield under water limiting conditions. PhD diss. Gainesville, Fla.: University of Florida
41. Liu, J., C. E. Goering, and L. Tian. 2001. Neural network for setting target corn yields. *Trans. ASAE* 44(3): 705-713. <https://elibrary.asabe.org/abstract.asp?aid=6097>
42. Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by backpropagating errors. *Nature* 323 (6088), 533–536. <https://doi.org/10.1038/323533a0>
43. Dayhoff, J. E. 1990. *Neural Network Architectures: An Introduction*. New York, N.Y.: Van Nostrand Reinhold.
44. Khoshnevisan, B.; Rafiee, S.; Omid, M.; Mousazadeh, H. Development of an intelligent system based on ANFIS for predicting wheat grain yield on the basis of energy inputs. *Inf. Process. Agric.* **2014**, *1*, 14–22. [CrossRef] <https://doi.org/10.1016/j.inpa.2014.04.001>
45. Khoshnevisan, B.; Rafiee, S.; Omid, M.; Mousazadeh, H. Prediction of potato yield based on energy inputs using multi-layer adaptive neuro-fuzzy inference system. *Measurement* **2014**, *47*, 521–530. [CrossRef] <http://dx.doi.org/10.1016/j.measurement.2013.09.020>
46. Amid, S.; Mesri Gundoshmian, T. Prediction of output energies for broiler production using linear regression, ANN (MLP, RBF), and ANFIS models. *Environ. Prog. Sustain. Energy* **2017**, *36*, 577–585. [CrossRef] <https://doi.org/10.1002/ep.12448>
47. Vivas, E.; Allende-Cid, H.; Salas, R. A Systematic Review of Statistical and Machine Learning Methods for Electrical Power Forecasting with Reported MAPE Score. *Entropy* **2020**, *22*, 1412. [CrossRef] <https://doi.org/10.3390/e22121412>
48. Wang, X.; Huang, J.; Feng, Q.; Yin, D. Winter Wheat Yield Prediction at County Level and Uncertainty Analysis in Main Wheat-Producing Regions of China with Deep Learning Approaches. *Remote Sens.* **2020**, *12*, 1744. [CrossRef] <https://doi.org/10.3390/rs12111744>
49. Zhao, Y.; Potgieter, A.B.; Zhang, M.; Wu, B.; Hammer, G.L. Predicting Wheat Yield at the Field Scale by Combining HighResolution Sentinel-2 Satellite Imagery and Crop Modelling. *Remote Sens.* **2020**, *12*, 1024. [CrossRef] <https://doi.org/10.3390/rs12061024>
50. Felipe Maldaner, L.; de Paula Corrêdo, L.; Fernanda Canata, T.; Paulo Molin, J. Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. *Comput. Electron. Agric.* **2021**, *181*, 105945. [CrossRef] <http://dx.doi.org/10.1016/j.compag.2020.105945>
51. Sharma, L.K.; Singh, T.N. Regression-based models for the prediction of unconfined compressive strength of artificially structured soil. *Eng. Comput.* **2018**, *34*, 175–186. [CrossRef] <https://link.springer.com/article/10.1007/s00366-017-0528-8>
52. Peng, J.; Kim, M.; Kim, Y.; Jo, M.; Kim, B.; Sung, K.; Lv, S. Constructing Italian ryegrass yield prediction model based on climatic data by locations in South Korea. *Grassl. Sci.* **2017**, *63*, 184–195. [CrossRef] <https://doi.org/10.1111/grs.12163>
53. Kim, S.; Kim, H. A new metric of absolute percentage error for intermittent demand forecasts. *Int. J. Forecast.* **2016**, *32*, 669–679. [CrossRef]

<https://econpapers.repec.org/scripts/redir.pf?u=https%3A%2F%2Fdoi.org%2F10.1016%2Fj.ijforecast.2015.12.003;h=repec:eee:intfor:v:32:y:2016:i:3:p:669-679>

54. Bhojani, S.H.; Bhatt, N. Wheat crop yield prediction using new activation functions in neural network. *Neural Comput. Appl.* **2020**, *32*, 13941–13951. [CrossRef] <https://doi.org/10.1007/s00521-020-04797-8>

55. Singh, R.; Umrao, R.K.; Ahmad, M.; Ansari, M.K.; Sharma, L.K.; Singh, T.N. Prediction of geomechanical parameters using soft computing and multiple regression approach. *Measurement* **2017**, *99*, 108–119. [CrossRef] <https://doi.org/10.1016/j.measurement.2016.12.023>

56. Chen, J.-F.; Do, Q.; Nguyen, T.; Doan, T. Forecasting Monthly Electricity Demands by Wavelet Neuro-Fuzzy System Optimized by Heuristic Algorithms. *Information* **2018**, *9*, 51. [CrossRef] <http://dx.doi.org/10.3390/info9030051>

57. Gandhi, N.; Petkar, O.; Armstrong, L.J. Rice crop yield prediction using artificial neural networks. In Proceedings of the 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, India, 15–16 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 105–110. <https://doi.org/10.1109/TIAR.2016.7801222>

58. Gandhi, N.; Armstrong, L.J.; Petkar, O.; Tripathy, A.K. Rice crop yield prediction in India using support vector machines. In Proceedings of the 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 13–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–5. <https://doi.org/10.1109/JCSSE.2016.7748856>

59. Schwalbert, R.A.; Amado, T.; Corassa, G.; Pott, L.P.; Prasad, P.V.V.; Ciampitti, I.A. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agric. For. Meteorol.* **2020**, *284*, 107886. [CrossRef] <https://doi.org/10.1016/j.agrformet.2019.107886>

60. Mishra, S.; Paygude, P.; Chaudhary, S.; Idate, S. Use of data mining in crop yield prediction. In Proceedings of the 2018 2<sup>nd</sup> International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 19–20 January 2018; pp. 796–802. <http://dx.doi.org/10.1109/ICISC.2018.8398908>

61. Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D., Pozza, L.E., Ugbaje, S.U., Jephcott, T.G., Paterson, S.E., Whelan, B.M., Bishop, T.F., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agric.* 1–15. <https://doi.org/10.1007/s11119-018-09628-4>

62. Tao, F.; Xiao, D.; Zhang, S.; Zhang, Z.; Rötter, R.P. Wheat yield benefited from increases in minimum temperature in the Huang-Huai-Hai Plain of China in the past three decades. *Agric. For. Meteorol.* **2017**, *239*, 1–14. <http://dx.doi.org/10.1016%2Fj.agrformet.2017.02.033>

63. Babkenov A.T., Babkenova S.A., Abdullayev K.K., Kairzhanov Y.K. Breeding Spring Soft Wheat for Productivity, Grain Quality, and Resistance to Adverse External Factors in Northern Kazakhstan // *Journal of Ecological Engineering*. — 2020. — Vol. 21, № 6. — P. 8–12. — DOI: 10.12911/22998993/123160.

64. Zholaman R., Fedorenko E., Savin T., Aituganov K., Shvidchenko V. Influence of the North Kazakhstan plains weather on the productivity of the spring soft wheat // *Scientific Horizons*. — 2022. — Vol. 25, № 10. — P. 53–61. — DOI: 10.48077/scihor.25(10).2022.53-61.
65. Karatayev M., Clarke M., Salnikov V., Bekseitova R., Nizamova M. Monitoring climate change, drought conditions and wheat production in Eurasia: the case study of Kazakhstan // *Heliyon*. — 2022. — Vol. 8, № 1. — Art. e08660. — DOI: 10.1016/j.heliyon.2021.e08660.
66. Nurpeisov D.N., Koishibaev M.K., Bekseitova R.T., Toleubayev K.M. Influence of weather conditions in the steppe zone of Northern Kazakhstan on the productivity of spring soft wheat // *Eurasian Agrotechnical Journal*. — 2023. — № 2(117). — DOI: 10.51452/kazatu.2023.2(117).1370.
67. Romanovska P., Schauburger B., Gornott C. Wheat yields in Kazakhstan can successfully be forecasted using a statistical crop model // *European Journal of Agronomy*. — 2023. — Vol. 147. — Art. 126843. — DOI: 10.1016/j.eja.2023.126843.
68. Romanovska P., Undorf S., Schauburger B., Duisenbekova A., Gornott C. Human-induced climate change has decreased wheat production in northern Kazakhstan // *Environmental Research: Climate*. — 2024. — Vol. 3, № 3. — Art. 031005. — DOI: 10.1088/2752-5295/ad53f7.
69. Morgounov A., Rosseeva L., Koyshibayev M. Leaf rust of spring wheat in Northern Kazakhstan and Siberia: incidence, virulence, and breeding for resistance // *Australian Journal of Agricultural Research*. — 2007. — Vol. 58, № 9. — P. 847–853. — DOI: 10.1071/AR07086.
70. Babkenova S.A., Babkenov A.T., Pakholkova E.V., Kanafin B.K. Pathogenic complexity of septoria spot disease of wheat in northern Kazakhstan // *Plant Science Today*. — 2020. — Vol. 7, № 4. — P. 601–606. — DOI: 10.14719/pst.2020.7.4.798.
71. Уалиева Р.М., Каверина М.М. Мониторинг видового состава вредителей яровой пшеницы на Северо-Востоке Казахстана // *Вестник Евразийского национального университета имени Л.Н. Гумилева. Серия: Биологические науки*. — 2023. — № 2(143). — С. 39–53. — DOI: 10.32523/2616-7034-2023-143-2-39-53.
72. Schauburger B., Jägermeyr J., Gornott C. A systematic review of local to regional yield forecasting approaches and frequently used data resources // *European Journal of Agronomy*. — 2020. — Vol. 120. — Art. 126153. — DOI: 10.1016/j.eja.2020.126153.
73. Shi W., Tao F., Zhang Z. A review on statistical models for identifying climate contributions to crop yields // *Journal of Geographical Sciences*. — 2013. — Vol. 23. — P. 567–576. — DOI: 10.1007/s11442-013-1029-3.
74. Lobell D.B., Burke M.B. On the use of statistical models to predict crop yield responses to climate change // *Agricultural and Forest Meteorology*. — 2010. — Vol. 150, № 11. — P. 1443–1452. — DOI: 10.1016/j.agrformet.2010.07.008.
75. Conradt T. Choosing multiple linear regressions for weather-based crop yield prediction with ABSOLUT v1.2 applied to the districts of Germany //

International Journal of Biometeorology. — 2022. — Vol. 66. — P. 2287–2300. — DOI: 10.1007/s00484-022-02356-5.

76. Lee B.H., Kenkel P., Brorsen B.W. Pre-harvest forecasting of county wheat yield and wheat quality using weather information // *Agricultural and Forest Meteorology*. — 2013. — Vol. 168. — P. 26–35. — DOI: 10.1016/j.agrformet.2012.08.010.

77. Bezabih G., Wale M., Satheesh N., Fanta S.W., Atlabachew M. Forecasting cereal crops production using time series analysis in Ethiopia // *Journal of the Saudi Society of Agricultural Sciences*. — 2023. — Vol. 22, № 8. — P. 546–559. — DOI: 10.1016/j.jssas.2023.07.001.

78. Singh K.N., Singh K.K., Kumar S., Panwar S., Gurung B. Forecasting crop yield through weather indices through LASSO // *The Indian Journal of Agricultural Sciences*. — 2019. — Vol. 89, № 3. — P. 540–544. — DOI: 10.56093/ijas.v89i3.87602.

79. van Klompenburg T., Kassahun A., Catal C. Crop yield prediction using machine learning: A systematic literature review // *Computers and Electronics in Agriculture*. — 2020. — Vol. 177. — Art. 105709. — DOI: 10.1016/j.compag.2020.105709.

80. Liakos K.G., Busato P., Moshou D., Pearson S., Bochtis D. Machine Learning in Agriculture: A Review // *Sensors*. — 2018. — Vol. 18, № 8. — Art. 2674. — DOI: 10.3390/s18082674.

81. Breiman L. Random Forests // *Machine Learning*. — 2001. — Vol. 45. — P. 5–32. — DOI: 10.1023/A:1010933404324.

82. Jeong J.H., Resop J.P., Mueller N.D., Fleisher D.H., Yun K., Butler E.E. et al. Random Forests for Global and Regional Crop Yield Predictions // *PLOS ONE*. — 2016. — Vol. 11, № 6. — Art. e0156571. — DOI: 10.1371/journal.pone.0156571.

83. Uribeetxebarria A., Castellón A., Aizpurua A. Optimizing Wheat Yield Prediction Integrating Data from Sentinel-1 and Sentinel-2 with CatBoost Algorithm // *Remote Sensing*. — 2023. — Vol. 15, № 6. — Art. 1640. — DOI: 10.3390/rs15061640.

84. Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators // *Neural Networks*. — 1989. — Vol. 2, № 5. — P. 359–366. — DOI: 10.1016/0893-6080(89)90020-8.

85. Khaki S., Wang L. Crop Yield Prediction Using Deep Neural Networks // *Frontiers in Plant Science*. — 2019. — Vol. 10. — Art. 621. — DOI: 10.3389/fpls.2019.00621.

86. Khaki S., Wang L., Archontoulis S.V. A CNN-RNN Framework for Crop Yield Prediction // *Frontiers in Plant Science*. — 2020. — Vol. 10. — Art. 1750. — DOI: 10.3389/fpls.2019.01750.

87. Shahhosseini M., Hu G., Huber I., Archontoulis S.V. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt // *Scientific Reports*. — 2021. — Vol. 11. — Art. 1606. — DOI: 10.1038/s41598-020-80820-1.

88. Lischeid G., Webber H., Sommer M., Nendel C., Ewert F. Machine learning in crop yield modelling: A powerful tool, but no surrogate for science //

Agricultural and Forest Meteorology. — 2022. — Vol. 312. — Art. 108698. — DOI: 10.1016/j.agrformet.2021.108698.

89. Sadenova M., Beisekenov N., Varbanov P.S., Pan T. Application of Machine Learning and Neural Networks to Predict the Yield of Cereals, Legumes, Oilseeds and Forage Crops in Kazakhstan // Agriculture. — 2023. — Vol. 13, № 6. — Art. 1195. — DOI: 10.3390/agriculture13061195.

90. Mimenbayeva A., Issakova G., Tanykpayeva B., Tursumbayeva A., Suleimenova R., Tulkibaev A. Applying Machine Learning for Analysis and Forecasting of Agricultural Crop Yields // Scientific Journal of Astana IT University. — 2024. — Vol. 17. — P. 28–42. — DOI: 10.37943/17LKYP9288.

91. Levenberg K. A method for the solution of certain non-linear problems in least squares // Quarterly of Applied Mathematics. — 1944. — Vol. 2, № 2. — P. 164–168. — DOI: 10.1090/qam/10666.

92. Marquardt D.W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters // Journal of the Society for Industrial and Applied Mathematics. — 1963. — Vol. 11, № 2. — P. 431–441. — DOI: 10.1137/0111030.

93. Srivastava, A.K. et al. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data // Scientific Reports. — 2022. — Vol. 12. — Art. 3097. — DOI: 10.1038/s41598-022-06249-w

94. Ashfaq, M. et al. Predicting wheat yield using deep learning and multi-source environmental data // Scientific Reports. — 2025. — DOI: 10.1038/s41598-025-11780-7

95. Zachow, M. et al. Wheat yield forecasts with seasonal climate models and long short-term memory networks // Computers and Electronics in Agriculture. — 2025. — Vol. 239. — Art. 110965. — DOI: 10.1016/j.compag.2025.110965

96. Ma, J. et al. Winter Wheat Yield Prediction and Influencing Factors Analysis Based on FourierGNN–Random Forest Combined Modeling // Agronomy. — 2025. — Vol. 15. — Art. 641. — DOI: 10.3390/agronomy15030641

97. Cheng, E. et al. Wheat yield estimation using remote sensing data based on machine learning approaches // Frontiers in Plant Science. — 2022. — Vol. 13. — Art. 1090970. — DOI: 10.3389/fpls.2022.1090970

98. Hornik, K., Stinchcombe, M., White, H. Multilayer feedforward networks are universal approximators // Neural Networks. — 1989. — Vol. 2, № 5. — P. 359–366. — DOI: 10.1016/0893-6080(89)90020-8

99. Zhao, J. et al. Assessing the combined effects of climatic factors on spring wheat phenophase and grain yield in Inner Mongolia, China // PLoS ONE. — 2015. — Vol. 10, № 5. — Art. e0127590. — DOI: 10.1371/journal.pone.0127590

100. Iqbal, N. et al. Analysis of Wheat-Yield Prediction Using Machine Learning Models under Climate Change Scenarios // Sustainability. — 2024. — Vol. 16. — Art. 6976. — DOI: 10.3390/su16166976

101. Khosravani Shariati, S.A., Abbasi, A. Machine learning-based winter wheat yield prediction using multisource data // Agricultural Water Management. — 2025. — Vol. 311. — Art. 109391. — DOI: 10.1016/j.agwat.2025.109391

102. Jawad, H.M. et al. Energy-Efficient Wireless Sensor Networks for Precision Agriculture: A Review // *Sensors*. — 2017. — Vol. 17. — Art. 1781. — DOI: 10.3390/s17081781
103. Bayih, A.Z. et al. Utilization of Internet of Things and Wireless Sensor Networks for Sustainable Smallholder Agriculture // *Sensors*. — 2022. — Vol. 22. — Art. 3273. — DOI: 10.3390/s22093273
104. Saha, S. et al. Precision agriculture for improving crop yield predictions: a literature review // *Frontiers in Agronomy*. — 2025. — Vol. 7. — Art. 1566201. — DOI: 10.3389/fagro.2025.1566201
105. Aslan, M.F. et al. Artificial Intelligence Techniques in Crop Yield Estimation Based on Sentinel-2 Data // *Sustainability*. — 2024. — Vol. 16. — Art. 8277. — DOI: 10.3390/su16198277
106. Hagan, M.T., Demuth, H.B., Beale, M.H., De Jesús, O. *Neural Network Design*. 2nd ed. — Oklahoma State University, 2014. — 1012 p. — ISBN 978-0-9717321-1-7
107. Marquardt, D.W. An algorithm for least-squares estimation of nonlinear parameters // *SIAM Journal on Applied Mathematics*. — 1963. — Vol. 11, № 2. — P. 431–441. — DOI: 10.1137/0111030
108. Levenberg, K. A method for the solution of certain non-linear problems in least squares // *Quarterly of Applied Mathematics*. — 1944. — Vol. 2. — P. 164–168. — DOI: 10.1090/qam/10666
109. MathWorks. *Deep Learning Toolbox User's Guide (R2023b)*. — Natick: The MathWorks, Inc., 2023. — URL: <https://www.mathworks.com/help/deeplearning/> (дата обращения: 15.03.2025)

ПРИЛОЖЕНИЕ А  
Патент на полезную модель

КАЗАҚСТАН РЕСПУБЛИКАСЫ      РЕСПУБЛИКА КАЗАХСТАН

REPUBLIC OF KAZAKHSTAN

**ПАТЕНТ  
PATENT**

№ 9091

**ПАЙДАЛЫ МОДЕЛЬГЕ / НА ПОЛЕЗНУЮ МОДЕЛЬ / FOR UTILITY MODEL**

 (21) 2024/0337.2

(22) 01.03.2024

(45) 03.05.2024

(54) Топырақты талдау үшін деректерді жинау және өңдеу жүйесі  
Система сбора и обработки данных для анализа почв  
Data acquisition and processing system for soil analysis

(73) Молдахметов Саят Сайранович (KZ); Аубакирова Гайля Фаритовна (KZ)  
Moldakhmetov Sayat Saiganovich (KZ); Aubakirova Gailya Faritovna (KZ)

(72) Молдахметов Саят Сайранович (KZ)      Moldakhmetov Sayat Saiganovich (KZ)  
Аубакирова Гайля Фаритовна (KZ)      Aubakirova Gailya Faritovna (KZ)  
Риттер Дмитрий Викторович (KZ)      Ritter Dmitriy Viktorovich (KZ)



ЭЦК қол қойылды  
Подписано ЭЦП  
Signed with EDS

Е. Оспанов  
Е. Оспанов  
Y. Ospanov

«Ұлттық зияткерлік меншік институты» РМҚ директоры  
Директор РГП «Национальный институт интеллектуальной собственности»  
Director of RSE «National institute of intellectual property»

Патентті күшінде ұстау ақысы уақытылы төленген жағдайда патенттің күші  
Қазақстан Республикасының бүкіл аумағында қолданылады.

Патентке пайдалы модельдің толық сипатта масы [www.kazpatent.kz](http://www.kazpatent.kz) ресми сайтында  
«Қазақстан Республикасының пайдалы модельдерінің мемлекеттік тізілімі» бөлімінде қолжетімді

\* \* \*

Действие патента распространяется на всю территорию Республики Казахстан  
при условии своевременной оплаты поддержания патента в силе.

Полное описание полезной модели к патенту доступно на официальном сайте [www.kazpatent.kz](http://www.kazpatent.kz)  
в разделе «Государственный реестр полезных моделей Республики Казахстан».

\* \* \*

Subject to timely payment for the maintenance of the patent in force  
the patent shall be effective on the entire territory of the Republic of Kazakhstan.

Full description of the patent for utility model are available on the official website [www.kazpatent.kz](http://www.kazpatent.kz)  
in the section «State Register of Utility Models of the Republic of Kazakhstan».



Қазақстан Республикасы Өділет министрлігінің  
«Ұлттық зияткерлік меншік институты» РМҚ  
Астана қаласы, Мәңгілік Ел даңғылы, ғимарат 57А

РГП «Национальный институт интеллектуальной собственности»  
Министерства юстиции Республики Казахстан  
Город Астана, проспект Мангилік Ел, здание 57А

«National Institute of Intellectual Property» RSE,  
Ministry of Justice of the Republic of Kazakhstan  
Astana, 57A Mangilik El Avenue

Тел./Tel.: +7 (7172) 62-15-15  
E-mail: [kazpatent@kazpatent.kz](mailto:kazpatent@kazpatent.kz)  
Website: [www.kazpatent.kz](http://www.kazpatent.kz)

## ПРИЛОЖЕНИЕ В

### Акт внедрения

Директор  
ТОО «Агрофирма  
«Мамлютская»  
*В. А. Грачев*  
В. А. Грачев  
«09» августа 2021 г.

#### АКТ

**о внедрении результатов научно-исследовательской работы в рамках диссертации Г.Ф. Аубакировой на тему «Разработка системы сбора и обработки данных для прогноза урожайности зерновых в СКО на основе искусственной нейронной сети»**

Мы, нижеподписавшиеся, составили настоящий акт о том, что система сбора и обработки данных для прогноза урожайности, разработанная в рамках диссертационной работы Аубакировой Гаили Фаритовны, использовалась для мониторинга урожайности на поле ТОО «Агрофирма «Мамлютская» с 6 мая по 9 августа 2021 г.

Объектом внедрения НИР выступает система, представляющая устройство сбора и передачи данных для построения прогнозной модели на основе искусственной нейронной сети, состоящая из одного стационарного модуля сбора данных и 4 переносных модуля.

В рамках процедуры внедрения результатов НИР были выполнены следующие работы. На открытом поле были установлены 4 переносных модуля, с помощью датчиков собирающих данные, которые используются для построения прогноза урожайности. Каждый модуль был установлен на мачту высотой от 0,8 до 1,2 м, что обеспечивает хорошую обзорность установки. Переносной модуль включает в себя 3 датчика влажности почвы, каждый из которых был погружен на глубину 1 м, 0,5 м и 0,1 м соответственно. Датчики температуры почвы и кислотности были погружены в почву на глубину 1 м. Датчик температуры воздуха установлен непосредственно на самом модуле.

Переносной модуль сбора данных, представляет собой электронное устройство, помещенное в прозрачный бокс, включает в себя:

- отладочный комплекс Arduino UNO R3 на базе микроконтроллера ATmega328 с платой расширения интерфейсов собственного производства и кнопочной панелью;

- 20-знаковый 4-строчный ЖКИ с I2C конвертером;

- фотоэлектрический преобразователь номинальной мощностью 5 Вт, модуль зарядки аккумуляторов на схеме TP4056 с функцией защиты от перезаряда и переразряда, 2 литий-ионных аккумуляторных батареи 18650 номинальным напряжением 3.7 В в батарейном отсеке;