Exploring Machine Learning Techniques for Symptom-Based Detection of Livestock Diseases

by

MAHIR AHMED NILOY
19101114
TANMAY BHOWMIK
19101465
JENNIFER ABEDIN
20301219
SYEDA JANNATUL FERDOUS
20301067
ISHRAT JAHAN
20301152

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfilment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
May 2024

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

MAHIR AHMED NILOY

19101114

TANMAY BHOWMIK

19101465

SYEDA JANNATUL FERDOUS

20301067

ISHRAT JAHAN

20301152

JENNIFER ABEDIN

20301219

Approval

The thesis titled "Exploring Machine Learning Techniques for Symptom-Based Detection of Livestock Diseases" submitted by

1. Mahir Ahmed Niloy (19101114)
2. Tanmay Bhowmik (19101465)
3. Jennifer Abedin (20301219)
4. Syeda Jannatul Ferdous (20301067)
5. Ishrat Jahan (20301152)

of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on May 21, 2024.

Examining Committee:

Supervisor:

Dr. Jannatun Noor

Assistant Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Program Coordinator:
(Member)

Dr. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Head of Department:

Dr. Sadia Hamid Kazi

Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Acknowledgement

We begin by expressing our gratitude to the Almighty Allah for preserving our physical and mental well-being and allowing us to execute our thesis work, ”Exploring Machine Learning Techniques for Symptom-Based Detection of Livestock Diseases,” seamlessly and within the designated time frame. We also acknowledge the divine endowment of knowledge, skills, and determination that empowered us to accomplish our task.

Our heartfelt appreciation goes out to our dedicated supervisor, Dr. Jannatun Noor, PhD, an Assistant Professor in Computer Science and Engineering, for her invaluable guidance, unwavering support, continuous motivation, and constructive suggestions throughout our thesis journey. Her enduring assistance has been instrumental in comprehensively developing our work and ideas.

The successful outcome of this study is not solely the result of individual toil, but rather a product of the collective efforts of numerous individuals. We extend our gratitude to our parents, whose unwavering hope and ongoing support have enabled us to pursue our career aspirations. Without their unwavering support, completing our thesis work would have been an uphill battle.

Abstract

Effective livestock monitoring ensures food security and sustainability in our rapidly growing world. However, proper cattle disease is still not taken seriously in our country. Even in the livestock industry, it has not become important yet. Very few livestock farms in Bangladesh collect data on their cattle, so gaining enough data is very tough. Most farm owners are not interested in collecting data; they fear the cost of IoT-based digital farms. Cost is a major concern for small farms as well. The proposed research aims to analyse the application of ML models in this specific sector of livestock management which is disease detection, by analysing various symptoms. Traditionally, Bangladeshi farms provide initial treatment to cattle based on symptoms. Most veterinary doctors in the village used these techniques as a tool for disease detection. We have worked with a dataset of about 43800 instances where almost 28 symptoms were used to detect a disease accurately. Advanced machine learning models such as Neural Network, Gradient boosting classifier, Decision tree classifier, Random forest, XGBoost, KNN etc. were used to determine possible diseases based on the collected symptoms. Overall, this research seeks to provide valuable insights and proper mitigation techniques into the livestock industry by analysing the impact of disease, as this will reduce mortality rates, fulfil the market demand for protein, and bring benefits to the dairy industry.

Keywords: Livestock Disease Detection, Disease Symptoms, Neural Network, Gradient Boosting Classifier, Ensemble Model.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviations that will be later used within the body of the document

DT	Decision Tree
EDA	Exploratory Data Analysis
GBM	Gradient Boosting Machine
IoT	Internet of Things
KNN	K-Nearest Neighbor
MLR	Multiple Linear Regression
NN	Neural Network
OLS	Ordinary Least Square
RF	Random Forest
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Background

The livestock industry is the foundation of global agriculture as it has served the world as an essential source of nutrition, earnings, and employment for many people and communities. Its importance is not only bounded by animal-based products but extends to many more sectors, such as food security, rural development, and economic sustainability. The demand for livestock products has been increasing rapidly due to the fast growth of the world population which is creating significant challenges. According to the FAO, livestock maintains about 1.3 billion people's livelihoods and ability to eat and be healthy. Also, it contributes 40% of the value of the world's agricultural output [27]. Therefore, this sector must maintain balance for our sustainability.

According to the information from the Department of Livestock Services, the livestock industry of Bangladesh contributes 1.47% of the country's GDP, and its GDP growth rate is 3.47%; about 20% of the nation's population works directly in the livestock industry.[11] Hence, we can observe the severity of the livestock diseases. Additionally, farmers are negligent about symptoms, which adds more trouble to controlling damage from livestock diseases. Furthermore, animal disease detection needs to initiate some tests, which are costly and time-consuming. Moreover, it is prone to errors frequently. On the other hand, machine learning models (ML) can work simultaneously and detect diseases faster than traditional methods. Thus, we came up with this research idea to detect disease at an early stage through machine learning. Our research will impact farmers' ability to detect disease so that they can take preventive measures. Lastly, if the diseases remain untreated, they will impact price volatility, disrupt supply chains, and reduce productivity. As a result, it will hamper the economic growth of a country along with the livelihoods of those dependent on this industry.

Figure 1.1: Livestock Farm

1.2 Motivation

The main objective of this research is to detect diseases so that farmers can take measures to prevent the spread of these diseases. Early identification of the diseases will enable prompt treatment and reduce economic losses for farmers. Also, it will maintain optimal livestock productivity by decreasing the mortality rate. Nowadays, people are facing problems due to the decrease in livestock and the effects of various deadly diseases, which are reducing the productivity of animals. This is causing profit losses as well as impacting market fluctuations. Our disease detection will help farmers confine affected animals, adopt suitable treatment plans, and take preventive action to protect the herd's general health when infections are discovered early. Besides, early illness identification in animals provides consumers with security and safety. Early identification stops diseases from spreading and guarantees that clean, high-quality animal products are produced. Our machine learning models, including Neural Networks, Gradient Boosting Classifier, Decision Tree Classifier, Random Forest, and XGBoost, are helpful because they take less time to execute and have good accuracy. To summarise, our study integrates disease detection to support farmers and deliver safeguards to them. Our method is effective because it uses modern machine-learning models and offers the potential for a more robust and sustainable agricultural system.

We focus on our research on making it easier to predict zoonotic diseases and providing effective support to farmers against these diseases so that they can be aware of the small conditions of affected animals. They can then administer medical procedures faster, reduce economic losses, and ensure improved livestock productivity. Additionally, early checkups of this disease are important to help stabilize the supply of surplus products such as milk, meat, and other agricultural products.

The journey to modern agroecology is poised for tragic challenges from the proportionate decline in animal numbers and the endemic risk of various devastating diseases. This disadvantage reduces animal productivity, causes financial loss to farmers, and increases market volatility. Our disease early checkup initiatives are designed to arm farmers to isolate affected animals, introduce specialized treatment plans, and ensure compliance with their hard work before the disease is suspected.

In addition, earlier detection of animal diseases provides important assurance and security to the original customer. We contribute to building consumer confidence in the agricultural sector by ensuring that the spread of disease is minimized and that clean, high-quality animal products are produced. This also improves restorative market stability and seeks to secure agricultural production.

We extensively use pioneering machine learning methods such as decision tree classifiers, Random Forests, and XGBoost models to optimize the disease detection process for speed and accuracy. These models do not deviate from particular results by providing rapid performance, which is integral to their invaluable preparation in unnecessary walls in our efforts to preserve and improve sustainable agriculture.

Overall, our research initiatives rely on pioneering disease detection methods to integrate with farmers, gain trust, and provide consumer safety. Using the power of modern machine learning models, we develop a robust and sustainable agricultural system that improves economic stability and creates improved security.

1.3 Research Questions

1. How will this research help the livestock industry locally and globally?

2. How will we achieve our final goal by machine learning?

3. How effective the machine learning models will be in detecting diseases based on symptoms??

4. How does the timely detection and treatment of livestock diseases influence the overall livestock industry?

1.4 Problem Statement

The livestock market has been facing serious problems due to different diseases. It is important to detect these diseases earlier to stabilize this sector. However, many factors are playing a significant role in worsening this situation. One big issue is the lack of good technology. It means diseases are not spotted early enough. This delay causes the diseases to spread more and raises the number of animals that die. Actually, the traditional ways of finding diseases are also a problem because they take a long time and cost a lot of money, which makes it tough for small farms to handle the situation. Additionally, sometimes these methods can be wrong. As a result, it leads to incorrect diagnosis. It can be challenging for small farms to bear the substantial costs of comprehensive disease testing. Our research focuses on symptom-based identification to quickly and accurately detect diseases. Focusing on observable symptoms, we can develop targeted plans to address their impact and ensure livestock's overall health and well-being.

1.5 Limitations

It is important to address and acknowledge several key limitations, although our research presents significant and promising results. The wide variety of livestock species and their production methods are complex, limiting the universal applicability of our findings. Our conclusions may need to be adapted for different areas of the livestock industry. It is crucial for our machine learning models to have up-to-date and accurate information on livestock diseases. Acquiring this data is challenging due to costs, consistency issues, availability, and reliability. Additionally, differences in reported information across regions can create uncertainty, affecting the robustness of our models. Furthermore, data collection is also complicated by the dynamic and widespread nature of the livestock market, which limits our models' effectiveness. Ensuring the models' relevance and durability requires continuous updates and adaptations to changing conditions. Scaling the models to cover the entire livestock industry may face logistical and financial challenges beyond our current focus. Our research provides valuable insights into machine learning for livestock disease detection, but addressing these obstacles is essential for ensuring our methods' relevance, feasibility, and durability in the dynamic livestock industry.

1.6 Research Contribution

Our research aims to enhance the early detection of diseases within the dairy sector. Basically, it focuses on key livestock species such as cows, goats, sheep, and buffalo. We targeted prevalent diseases, including anthrax, pneumonia, foot and mouth disease, blackleg, and lumpy virus. Actually, we want to predict diseases at their earliest stages by closely monitoring visible symptoms exhibited by these animals. This profound approach allows us to reduce the harmful effects of diseases spreading among animals quickly. If we reduce mortality rates and minimize the severity of these diseases, our efforts will improve animal welfare and sustainability in the dairy industry. Our methodology comprehensively considers various factors such as symptoms, age, and temperature. In this field, we tried to collect data to predict diseases accurately. However, we cannot gain better results due to the shortage of proper data. We gained the highest accuracy using a neural network model in our default dataset[31], but the accuracy was unsatisfactory. Due to some common symptoms, our machine-learning model could not identify them correctly. Then, we used our second dataset [36], which was basically used to extract more symptoms of these diseases, and after merging these two datasets, we created a custom dataset for our research work. This customized dataset was implemented by machine learning models, and this time, the result is better than the previous one as farms can identify diseases more accurately. By adding some extra variables, we contributed to enhancing the overall detection process's performance. This approach to disease prediction allows stakeholders to make decisions at the right time. It also helps them take appropriate measures to safeguard the industry's stability. Therefore, our work aims to make the dairy industry stronger by promoting better ways to handle diseases and ensuring a good balance between disease prevention and treatment efforts.

1.7 Thesis Organization

1. In Chapter 1, we have basically highlighted the background information on Livestock Disease and the current monitoring system of this Industry. We have included the problem we figured out in this industry and set our goal for this research work.

2. In Chapter 2, we have included literature in this specific field with some related works.

3. In Chapter 3, we have briefly discussed our dataset and a problem we find in the dataset. Then, we picked our second dataset and created our custom dataset by merging those two datasets. We have tried to analyze our data and provide information regarding it.

4. In Chapter 4, carries the workflow we followed throughout our research work.

5. In Chapter 5, we discuss the models we utilized. In this chapter, we tried to provide basic information about models and included diagrams of those models to help readers understand them.

6. In Chapter 6, has the overall information of our experiments and output we find through it. We have discussed our overall output in phases here. We also discussed evaluation metrics, and then we created a comparison of the outputs we got from various models. Lastly, we end this chapter by including the overall findings.

7. In Chapter 7, is the segment where we have discussed our significance. We have also introduced our application prototype in this segment. However, throughout the research work, we have faced many challenges. All of them are included in this chapter.

8. In chapter 8, we have discussed our future plan and how we can achieve it by solving challenges. Besides, we have concluded by giving an overall idea of how this research will contribute to the livestock industry.

Chapter 2

Literature Review

The livestock industry in Bangladesh has emerged as a vital component of the country's economy, playing a crucial role in ensuring food security, economic growth, and social empowerment. As outlined in the preceding introduction, this comprehensive study delves into the various facets of this burgeoning sector. This literature review section expands on the insights presented, drawing from an extensive analysis of several scholarly papers clustered into distinct categories: Cost, Health and Disease, Technology, and Environment.

2.1 Livestock Production and Cost:

The study [27] states that the primary goal is to work on the welfare of pigs in India by focusing on three main variables: cost, welfare, and production, as in most cases, those animals are produced abnormally and kept in terrible conditions.

Regression is a perfect methodology for identifying significant correlations between variables. In this paper, they used Multiple Linear Regression (MLR) as they had one response variable and multiple explanatory variables.
The equation used here:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{2.1}$$

$\hat{Y}$ = response variable
X_1, X_2 = explanatory variables
β = regression coefficient

However, this model was invalid because the squared value was relatively low (0.39), and there was no linear relationship between the three variables. Due to state government restrictions, getting data or statistics on laws and regulations in some places was complex. As a result, it wasn't easy to create a proper connection or generate variable relationships for the whole country's livestock industry.

Regarding the welfare data, two welfare-related factors have been considered: the amount of space allocated for each animal and the state-wide vaccination of all animals. These two factors can be essential to ensuring animal welfare and can benefit the livestock industry.

The research [26] is a significant undertaking motivated by the central role of agriculture, particularly livestock farming, in Pakistan's economy [26]. It is important to understand the role of livestock in our economy and agriculture. The researchers examined the air quality in the districts of Kasur and Lahore on ten distinct farms. They tried to find out if the air quality impacts the outbreaks of respiratory diseases. Moreover, this research examined a wide range of variables, including- temperature, humidity, carbon dioxide (CO2) levels, and many more. They conducted this experiment in the summer and winter seasons. The study found out the potential risks for animals and workers due to poor air quality, dust and microbes.

Despite some limitations, such as the study's regional and seasonal focus, the findings provide valuable insights into the current air quality status of livestock farms. The research stated that air quality is important to maintain by checking farm activities and maintaining the PM threshold level [26]. Therefore, the study focuses on the importance of preserving the air quality to control disease outbreaks and also to achieve environmental sustainability [26].

The research [18]focuses on retaining good quality pellet feed by using machine learning models. These feeds play a crucial role in increasing animal production, as well as they can be kept by using minimum storage. The researchers collected many elements that are important to enhance the quality of the feed by visiting mills. To get an accurate outcome for the animal feed pellet durability index (PDI), they used several types of regression analysis. Among the algorithms considered, the Support Vector Regression (SVR) method is identified as the top performer, with SR and Linear Support Vector Regression (LSVR) following closely behind. The determining criteria of significance included the production process, bakery by-product and wheat inclusion levels, ambient temperature, and pellet quality. Moreover, the algorithms consistently identify the average temperature and the presence of a bakery as the most prominent characteristics. The study acknowledged many limitations, including the absence of data regarding crucial aspects such as the particle size of ingredients and the makeup of nutrients. Furthermore, the models tested the identical dataset that was utilized for training purposes, and it was suggested that future evaluations be conducted on new data in order to evaluate the models' ability to generalize. In conclusion, the main outcome of this study indicates that the application of machine learning methodologies can enhance the predictive comprehension of pellet quality within the context of commercial feed production. This advancement holds the potential to enhance quality management practices and operational effectiveness.

Moreover, the algorithms consistently identify the average temperature and bakery's presence as the most prominent characteristics. The study acknowledged many limitations, including the absence of data regarding crucial aspects such as the ingredients' particle size and the nutrients' makeup. Furthermore, the models tested the identical dataset utilized for training purposes, and it was suggested that utilised evaluations be conducted on new data to evaluate the models' ability to generalize. In conclusion, the primary outcome of this study is that the application of machine learning methodologies can enhance the predictive comprehension of pellet quality within the context of commercial feed production. This advancement holds the potential to enhance quality management practices and operational effectiveness.

In this research[4], the intriguing nexus between microbiome data and machine learning in the context of swine farming. The benefit of swine production to the agricultural sector is acknowledged at the outset of the report, as is the increased interest in using microbiome data to improve productivity and meat quality. Using machine learning to forecast growth and carcass qualities is a revolutionary strategy with great promise to increase the effectiveness of swine farming.

One of this work's advantages is its thorough approach. To create production models, the authors gathered microbiome data from pigs and applied a range of machine-learning methods. This method thoroughly assesses how well various algorithms predict growth and carcass features. It is admirable that model comparison and assessment measures have been included since they increase openness and make it easier to repeat the study.

The research [4] also explores potential biological pathways that might be responsible for the relationships between growth features and the swine microbiome that have been discovered. However, the requirement for bigger and more varied datasets is a possible restriction for the research. In conclusion, the methodology rigour of the study and the addition of biological insights increase its worth. The publication offers insightful contributions and provides a robust framework for further investigation into microbiome research in agriculture, which is still in its early stages.

The above discussions cover various aspects of livestock production, sustainability, and innovations in the livestock business that are relevant to the topic we are working on. They emphasize the significance of addressing beef cattle production difficulties, taking environmental impact into account, and coming up with creative solutions. In order to ensure the resilience and sustainability of the livestock industry, these summaries place a strong emphasis on the need for sustainable practices, cutting-edge technologies, and interdisciplinary efforts. The analysis of the livestock market is crucial for ensuring food security [17]. As the livestock industry plays an important part in the economy, people should focus on this sector more for better outcomes.

In the research [23], the obstacles of the beef cattle sector and the required innovation of sustainability have been analyzed briefly. To satisfy the expected rise in global population, especially in central Africa, approximately 120 million kg of protein are needed. Particularly, the methane emissions from ruminant animals, as an impact of beef production, create a substantial concern. Struggling a balance between increasing protein output and maintaining the sustainability of the environment is critical. Animals like cows provide vital nutrients, such as amino acids from the non-edible plant's iron and zinc, which we cannot directly take from them. They turn them into important substances for our body, neutrinos. An eco-friendly cattle farm requires well-organized grazing methods that will improve soil health, soak up carbon, and help fight the effects of climate change. The management of forages can be improved, and livestock production efficiency can be increased by integrating environmental biology and technology such as remote sensing and machine learning. Finally, this study emphasizes the need to tackle the issues of sustainable beef cattle production in the context of global protein demands. The report [23] emphasizes the importance of interdisciplinary efforts and ensures the beef cattle industry's resilience and sustainability while addressing global development and environmental changes.

The paper [16] states that the implementation of machine learning techniques to enhanced reproductive management in dairy cattle is projected to yield notable enhancements in performance, profitability, and sustainability within the dairy industry, while also optimizing the management of dairy cattle. The primary focus of this study revolves around the concept of targeted reproductive management (TRM). TRM involves the identification of specific groups of cows with varying levels of reproductive potential, utilizing a range of biological, managerial and performance data sources. The applications of Targeted Reproductive Management (TRM) encompass strategic pregnancy timing while nursing, the use of sexed semen or embryo transfer to enhance the value of offspring, the implementation of targeted hormonal therapy, and the identification of optimal timing for artificial insemination. They are tracking a variety of cow-related data using sensors and machine learning algorithms. In order to help farmers better manage their reproductive tactics, these sensors monitor things like when cows are ready to mate. Ensuring the health and happiness of those cows while optimizing herd performance is the main goal. Moreover, there are more elements that influence their ability to reproduce, such as their biological makeup, management style, and surroundings. Finally, in dairy farming, Targeted Reproductive Management (TRM) is used to improve cow reproduction and streamline the entire process. Their strategies are customized to target particular groups of cows that share similar characteristics or predicted behaviours. They are able to maintain efficiency and produce greater outcomes as a result.

This study [13] stated that they wanted to have a clear understanding of how rural livestock farmers can adopt ICT tools in their businesses. Thus, this modern technology helps to improve livestock productivity and management. The goal of the project was to equip these farmers with the information and tools they needed to increase their output and have a greater economic effect. In this study[13] they wanted to find out which ICT devices are commonly used by rural livestock farmers in the region. They discovered that the highest % of people were using mobile phones, which was 91.3%. It only serves to highlight the significant influence that these technologies have on rural agricultural communities. As technology was already being used, researchers realized that reductions in production costs, management of livestock health, and prediction of disease outbreaks could be prevented through technological tools. However, there are some limitations to this research. The limitations include the expense of technology and poor technological infrastructure. Additionally, they used a quantitative method instead of a specific model. To conclude, the study emphasizes how much potential ICT tools have to transform Nigeria's rural livestock production. There is a huge potential for technology-driven agricultural development, as seen by the expanding use of mobile phones and the awareness of ICT's beneficial effects on livestock management and productivity. Developing policies and interventions that fully utilize the trans-formative power of ICT in rural agriculture will be necessary going forward, as will addressing adoption barriers and conducting long-term research to evaluate the long-term effects.

This research paper[12] states that there is a critical need for livestock monitoring in emerging countries like Bangladesh, where the livestock sector significantly contributes to the national economy but faces challenges due to poor maintenance and health oversight. They reference prior research, such as Y. P. Pratama et al.'s smart collar device (2019) for cattle health monitoring and P. Khatate et al.'s work on wearable intelligent health monitoring systems (2018), which utilized sensors like DS18B20 and flex sensors. The authors also acknowledge the importance of wireless sensor networks in livestock monitoring, citing Handcock et al.'s use of these networks for animal behaviour and environmental interaction monitoring (2009). Machine learning's role in livestock farming is emphasized, particularly in weight measurement using both direct and indirect techniques. The proposed methodology in this study involves IoT-based sensors like DS18B20, DHT11, and ESP8266 for data collection and communication. The dataset comprises readings of various parameters, with machine learning algorithms like Decision Tree Classifier, Support Vector Machine, and Multi-Layer Perceptron employed for classification and prediction. Experimental results highlight the Support Vector Machine as the most accurate model. The discussion underscores the importance of user-friendly and cost-efficient systems for rural farmers and suggests that their IoT and machine learning-based systems can improve livestock management in Bangladesh. Overall, the paper offers a comprehensive view of IoT and machine learning applications in livestock monitoring, referencing relevant prior work, presenting a detailed methodology, and emphasizing the potential benefits for both farmers and the economy.

The study[21] states that the need to use AI and UAVs in Livestock is discussed in order to enhance control and reap financial rewards. Livestock is an essential worldwide business, and smart livestock is growing with a market growth forecast of \$43.37 billion by 2030 from \$18.12 billion in 2021. Standard farm security requires a lot of labour and may not cover the entire farm, leaving it open to fraud and other problems. In order to decrease human labour, save time, and protect agricultural properties. The research emphasizes the necessity of real-time computerized field monitoring. Drone technologies will have a 57% market increase in the business sector from 2021 to 2028, with the incorporation of drones into livestock being a main emphasis. The research assesses different deep learning models. Such as Yolov7, RCNN, and SSD, to overcome the limitations of existing target recognition and tracking techniques. The report closes with a systematic strategy that includes summarizing relevant work, going into depth about data preparation, explaining the approach that was chosen, assembling models providing the results and the case study, analyzing the findings in relation to previous research, and defending data confidentiality. In conclusion, the research emphasizes the rising significance of automated livestock supported by AI and UAV technologies in improving livestock security and financial gains.

2.2 Livestock Health

The paper[29] states that the drawbacks of traditional livestock farming techniques are that they rely on outdated controls and weight-based sorting, which create inefficiency and disease risks. To create a deep learning-based sorting system, this study uses a dataset that has been enhanced with essential metrics and breeding photos gathered over a 24-month. It makes use of a Kalman filter-based strategy to improve sorting precision and a residual neural network (ResNet) for tracking cattle weight. In various situations, the Wasserstein Generative Adversarial Nets (WGAN) technique improves image identification. The technology enhances disease surveillance by identifying potentially unwell livestock by identifying unusual body traits.
As a consequence of the experiments, it was possible to identify sick pigs with a success rate of 98%, identify animals with an accuracy of 89%, recognize obscured images with a success rate of 32%, and accurate livestock with a success rate of 95%. Through efficiency in pig farming and cost-effective disease surveillance, this technology has the potential to transform the management of animals. In summary, these results highlight how deep learning-based technologies have the potential to transform livestock management and provide a promising answer to the problems that plague contemporary livestock, like pig farming.

This paper [30] focuses on the economic effects, risk factors, and prevalence of infectious diseases, primarily highlighting cattle. The data for this paper was obtained through various studies, investigations, and surveys. For example, the prevalence rate of brucellosis, contagious bovine pleuropneumonia(CBPP), foot and mouth disease(FMD), and peste des petits ruminants(PPR) were determined through serological tests, postmortem examinations and techniques like competitive ELISA(cELISA) and latex agglutination test. Both qualitative and quantitative data were used for this research. Among all the diseases, peste des petits ruminants(PPR) has the highest morbidity and infected flocks may experience a 90% mortality rate. For this purpose, this study recommends widespread education and awareness to reduce farmer's and herders' ignorance of nature and the spread of diseases. Besides, a flexible financing system is also advised to assist farmers so that they can easily access livestock healthcare. Furthermore, it emphasizes the necessity of control measures by authorities to emphasise the uncontrolled spread of diseases. Finally, further research can be done by examining how these diseases impact trade and economy and evaluating the effects on the domestic and international livestock trade.

The study [22] states that among all the animal diseases, FMD and diarrhea were the most common diseases that occurred among animals. These diseases follow a pattern that is simply dependent on the animal's age and breed. For example, internal parasite infestation is found commonly in young animals, whereas anestrus is more dominant in aged cattle. Moreover, diarrhea and internal parasite infestations are found to be primarily common based on their breed. To sum up, the researchers focused on the necessity of militarization to reduce the economic burden on farmers. According to [22], FMD and diarrhoea were the most prevalent diseases among 755 cattle and goats, respectively. Moreover, specific patterns are observed based on age and breed, underscoring variances in disease prevalence. For instance, anestrus is widespread in adult cattle, but internal parasite infestation is significant in the young. The study also focuses on specific cattle breeds, detailing prevalence rates for diseases like pneumonia and mastitis. Diarrhoea and internal parasite infestation are prominent in goats, with variations observed by age and breed. Finally, the research highlights the economic significance of ruminants and emphasizes the necessity for comprehensive disease management emphasises.

The study [14] states that there is a need for accurate and non-invasive methods to monitor animal health, behaviour, and identification on farms[14]. They explored various studies involving biometric data collection, such as heart rate (HR), respiratory rate (RR), and temperature measurements, utilizing contactless methods to reduce stress on animals. [14] employed webcams and the photoplethysmography (PPG) principle to assess pig HR with an accuracy of 0.80 (R^2). Similarly, Jorquera-Chavez et al. (2020) utilized integrated cameras to measure temperature, HR, and RR in pigs, reporting correlation coefficients ranging from 0.61 to 0.66 for HR and RR compared to manual measurements. These non-invasive techniques offer promising alternatives to traditional invasive methods.

The research [14] focuses on the monitoring of animal health and behaviour by employing various techniques. They revealed that the measurement of heart rate (HR), and respiratory rate (RR) should be contact-free to minimize the pressure on the animals. For example, the pig heart rate was acquired through webcams and photoplethysmography (PPG) to view the result. Furthermore, cameras were used to determine heart rate, and temperature in pigs and showed a significant result varying from 0.61 to 0.66. These contactless methods can be effective in determining biometric data for animals. This research also states the necessity of accurate health checkups and keeping disease track records. The authors highlighted the emergence of contactless biometric techniques, particularly in cattle recognition. These techniques included muzzle pattern identification, face recognition, and body recognition. Notable studies demonstrated the potential of deep learning models to recognize individual cattle based on these features, with varying accuracies [14]. Additionally, the authors discussed AI applications for assessing animal emotional responses, though they stressed the importance of objective measurements like hormonal data.

Regarding AI implementation, the review emphasized the need for practical deployment of AI solutions on farms. It cited instances where AI models remained in academic settings without real-world application. The authors recommended multidisciplinary teams comprising animal science, data analysis, and AI experts. They also advocated for starting with simple problems and leveraging historical data for AI model development [14]. Data quality and security were highlighted as fundamental concerns, suggesting the potential use of blockchain technology for secure data management

In conclusion, this review sheds light on the promising integration of biometrics and AI in livestock farming for health monitoring, identification, and welfare assessment[14]. It underscored the need for practical deployment, collaboration among experts, and secure data management for successful innovations in this field (Robinson et al., 2019). Future work should focus on bridging the gap between academic research and real-world applications to benefit both animal welfare and the livestock industry

This paper [19] states the effects of genetic selection for high egg output and large breast meat on the immune systems of domestic chickens, advancing veterinary epidemiology, livestock vaccine management, and economics. It underlines the impact of climate change on disease onset and draws attention to how susceptible modern poultry is to ailments like Newcastle disease and Avian Influenza because of regulated surroundings. Transboundary disease transmission is now more likely as a result of globalization. In order to increase poultry immunity, the study emphasizes the value of vaccination in illness prevention and productivity development. Weakened or inactivated viruses or proteins are used in this process. It underlines the necessity of individualized immunization schedules depending on a number of criteria. To conclude, this study offers insightful information on the poultry business's difficulties, emphasizing the importance of vaccination in reducing these difficulties and fostering sustainable chicken production, making it an essential tool for those involved in the sector.

The study[20] states the significance of enhancing livestock management efficiency by employing strategies such as feed formulation, diet optimization, livestock breeding, genetic optimization, and vaccine schedules. Feed formulation and diet optimization processes aid in providing animals with a nutritionally balanced diet while minimizing costs. Genetic optimization involves considering various elements, including genetic potential, heritability, and economic values associated with particular features. For instance, genetic algorithms or alternative optimization techniques might be employed to acquire the needed characteristics to attain such outcomes. This study examines disease transmission patterns, vaccine effectiveness, and cost restrictions in the context of disease control and vaccination scheduling. Optimization techniques can be employed to identify the most optimal schedule that simultaneously enhances protection and reduces expenses. The paper has certain limitations. The study briefly discusses data gathering as a component of the technique but lacks further details regarding the specific sources of data or the strategies employed to mitigate data quality concerns, which often arise in practical contexts. The discourse surrounding data concerns and potential solutions holds significant importance. Additionally, this report presents a broad overview of the findings, overlooking any specific instances in disease control, such as the prediction of disease outbreaks.

The research paper[5] states that the financial advantages of diagnostic testing for livestock, concentration on anaplasmosis in livestock. The expensive livestock illness anaplasmosis continues to spread worldwide, including in the United States.

The conventional control methods include immunization, antibiotics, post-infection treatments, and feed additives. The ideal mixture of disease control strategies is undetermined due to the anaplasmosis in livestock's highly contagious and frequently asymptotic beginning. The cost of implementing various anaplasmosis control choices for a representative cow-calf producer in the United States is calculated in the article to address this lack of clarity. The researchers include early detection using diagnostic testing in the current analysis as an extension. The findings imply that the best herd management method combines diagnostic tests with preventive medications. However, suboptimal control systems may be successful due to extra factors. The research presented here supports a more precise estimation of the anaplasmosis burden while offering an early investigation of the economic viability of diagnostic procedures. It is a vital instrument for anybody looking to improve decision-making around animal immunization through fast diagnostic procedures.

This study [6] focuses on establishing a stable decision-making framework for investors and livestock farms to choose the most suitable cattle breeds based on their geographical region. Additionally, the paper conducts a regression trend analysis of cattle categorized by their original breeds. The paper [6] implemented various machine learning models, including Multiple Linear Regression, Ordinary Least Square, Support Vector Machine, and Decision Tree learning, to generate accurate and optimized outcomes. The primary objective is to create and compare these models, to provide an effective plan to reduce protein demand in Bangladesh and contribute to lowering the unemployment rate. A significant challenge faced in this study [6] was the limited availability of data. Time series analysis typically requires data spanning a minimum of 25-30 intervals, but none of the farms had monthly-based data. The findings highlight the importance of motivating farms to collect monthly data, enabling the establishment of a robust monitoring system for the livestock industry. The research [6] indicates promising prospects for the future of Bangladesh's livestock industry. The utilization of advanced machine learning models offers a path to enhance decision-making for investors and farms, leading to the selection of the most suitable cattle breeds based on geographical regions. This has the potential to boost productivity and reduce protein demand in the country. Addressing data limitations through improved data collection methods, particularly monthly-based monitoring, can enhance decision-making and forecast disease outbreaks, resulting in cost savings for animal treatment. Embracing technology and data-driven insights will increase efficiency, reduce unemployment, and promote sustainable growth, solidifying the livestock industry's pivotal role in Bangladesh's economy and food security.

This research paper [33] states the application of machine learning techniques to predict agricultural prices, with a specific emphasis on the pork market. The statement highlights the importance of implementing automated decision support systems to rectify discrepancies in manual price determination. An extensive assessment of various machine learning models, such as Support Vector Machines (SVM), Random Forest, Ridge Regression, and Extremely Randomized Trees, is undertaken to determine the most effective methodology. This study makes a contribution by analyzing market data, taking into account temporal lags, and conducting a comparison of different models. These findings offer significant insights for practitioners interested in implementing machine learning techniques to make educated decisions in agricultural markets. In addition, this study also presents a prototype for the automation of weekly price projections, thereby showcasing the practical implications of employing these methodologies. The Ridge model continually demonstrated superior performance compared to other machine learning models. The selection of the data source, whether public or subscription-based, and the model's susceptibility to variations in data timing, were significant factors that influenced the predicted accuracy.

Consequently, an external intermediary called 'Lonjas' has been incorporated, utilizing a high-precision regression model. The model's predictions have consistently shown minimal disparities, often not exceeding 0.02 euros compared to the actual reference prices. This equates to an error rate of approximately 1.2%. Hence, this study aims to examine the influence of temporal lags in acquiring past data on the efficacy of a machine learning-based pork price forecasting system.

2.3 Impacts of Cattle Disease

The paper[1] states that assessing the profitability of livestock marketing, particularly goats and livestock, in the Gamawa Local Government Area of Bauchi State in 2012. Using structured questionnaires and a combination of purposive and straightforward random selection techniques, information was gathered from 120 livestock traders. Multiple regression analysis, qualitative data collection, and agricultural budgeting were used in the analysis. The survey found that goat and livestock marketers had average ages of 46, 40,11, and 13 years of marketing experience, and 9 and 13 people in their households, respectively. It was discovered that the marketing margins and returns on investment for goats and livestock, respectively, were N4,016.66 and N38,816.6, and N0.14 and N0.26.These numbers provide information on the regional livestock marketing industry's financial elements. The cost of goat purchase, medicine, and labour were significant factors determining goat marketing's profitability, whereas the cost of cow acquisition, feeding, medication, and labour affected livestock marketing. At various p-values, these variables showed differing degrees of significance. Poor financial facilities, a lack of market intelligence, excessive transportation expenses, and poor living facilities were among the major issues the survey outlined for livestock traders. It was discovered that these limitations affected the region's livestock marketing system's effectiveness and profitability. The financial implications of livestock marketing in Gamawa, the Local Government Area of Bauchi State, was the focus of the current research's analysis. While taking into account a variety of socio-socioeconomic meters, it evaluated the profitability of marketing goats and livestock. The research emphasized the value of elements that determine profitability and identified significant limitations in the cattle selling system. For stakeholders and politicians hoping to increase regional livestock selling, the industry's economic viability offers useful information.

This study [30] gives insight into the financial implications of prominent diseases, including mastitis, brucellosis, ketosis, hand-foot-and-mouth disease (FMD), and many more within a specified milk shed region, involving 184 livestock owners. Using a methodical survey, thorough data was acquired on disease occurrence, treatment timings, household situations, losses, treatment costs, and the historical background of these diseases, focused on 126 cows and 58 buffaloes. The Chi-square test was employed to evaluate economic losses, demonstrating a major impact on milk output and increased treatment costs. The findings underline a maintaining lack of awareness among livestock owners concerning essential elements such as diet, housing, healthcare, and hygiene for the widespread occurrence of these diseases. Moreover, brucellosis emerged as a key contributor to milk production losses, with higher morbidity rates found in buffaloes compared to cows. According to Patel, Komal. (2023), the economic impact of brucellosis was significant, with a loss of 1806.54 rupees per animal, mainly affecting milk production, which experienced a loss of 1317.20 rupees. Lastly, the study proposes preventive actions, including the deployment of a thorough surveillance program and the development of preventive methods to reduce economic losses.

The research [9] provides an in-depth analysis of the economic losses and broader effects resulting from the Lumpy Skin Disease(LSD), Sheep Pox(SP), and Goat Pox(GP) epidemics on small-scale farmers in Bauchi State, Nigeria. The data was acquired through structured interviews with 99 farmers impacted by lumpy skin disease, sheep pox, and goat pox epidemics. The parameters comprised treatment methods, expenses, adaptive techniques, production specifics, and financial damages, offering a thorough evaluation of the influence on the subsistence farmers' means of subsistence. The degree of correlation between the binary results of the two groups was established using chi-squared testing. Moreover, the production system has an impact on transhumance farmers since they suffer far greater losses. Additionally, the research estimates a 65% decrease in milk production in calves during the acute phase of LSD, and it quantifies both short- and long-term effects on a variety of production indices. Furthermore, the necessity for focused interventions is highlighted by the discovery that livestock markets could serve as major centres for the spread of viruses. The lack of vaccination availability in Nigeria highlights how urgent it is to put into place efficient disease control measures. As a whole, the study offers insightful information about the complexities of LSD, SP, and GP in the context of subsistence farming, highlighting the financial losses that occur right away, possible long-term effects, and important areas that should be addressed to reduce the spread of the disease and its negative effects on farmers' livelihoods.

The paper [2] discusses a comprehensive analysis of the milk production and marketing impact of Horro Guduru Wollega Zone, Western Ethiopia. This paper explores the dynamics, challenges, and opportunities within the Horror District. This study likely investigates aspects such as dairy farming practices, socio-economic factors influencing production, and the efficiency of marketing systems. The paper contributes valuable insights into the local dairy industry, potentially shedding light on the livelihoods of the communities involved and the broader economic implications. The standing of the system aids in identifying areas for improvement and potential interventions to enhance both the production and marketing aspects of the dairy. The findings may have practical implications for policymakers, practitioners, and stakeholders involved in the development of sustainable dairy practices in the region. Overall, this paper serves as a crucial resource for those seeking a little understanding of the milk production and marketing system in the Horror District.

This paper [24] focuses on the economic effects, risk factors, and prevalence of infectious diseases primarily highlighting cattle. The data for this paper was obtained through various studies, investigations, and surveys. For example, the prevalence rate of brucellosis, contagious bovine pleuropneumonia(CBPP), foot and mouth disease(FMD), and peste des petits ruminants(PPR) were determined through serological tests, postmortem examinations, and techniques like competitive ELISA(cELISA) and latex agglutination test. Both qualitative and quantitative data were used for this research. Among all the diseases, peste des petits ruminants(PPR) have the highest morbidity, and infected flocks may experience a 90% mortality rate. For this purpose, this study recommends widespread education and awareness to reduce farmer's and herders' ignorance of nature and the spread of diseases. Besides, a flexible financing system is also advised to assist farmers so that they can easily access livestock healthcare. Furthermore, it emphasizes the necessity of control measures by authorities to emphasise the uncontrolled spread of diseases. Finally, further research can be done by examining how these diseases impact trade and economy and evaluating the effects on the domestic and international livestock trade.

This research paper[32] states that locating feed-efficient livestock in the beef sector. Costly individual feed intake data are needed for conventional techniques like residual feed intake (RFI). To get around this, the study uses machine learning (ML) techniques with RFI readings and genotypic data from 4,057 beef cattle. From the 50 K Illumina Bovine panel, three approaches identify the best subsets of Single Nucleotide Polymorphisms (SNPs). Even with only a few SNPs, eleven ML algorithms accurately divide cattle into feed-efficient and inefficient groups. This strategy improves the economic sustainability of the beef business by lowering the cost and complexity of feed efficiency analyses. In conclusion, the study demonstrates that ML can accurately predict feed efficiency in cattle, providing a more affordable alternative to conventional RFI measures and enhancing the profitability and competitiveness of the beef production industry.

2.4 Related Work

Cattle health monitoring is gaining popularity daily since it has a vital impact on both individual and country economies. There are several ways of monitoring a cattle farm, but sustainability and affordability are the major concerns. Many proposed works exist, but all of them have some vital challenges that are stopping execution.

Previous works related to animal disease detection referred to the Internet of Things (IoT) and livestock monitoring. The research paper [12] stated the necessity of livestock monitoring by using sensors. These sensors are costly and need to be monitored briefly. Therefore, the small farms could not afford it. Besides the process is time-consuming. Moreover, the farmers are not well informed about this technology. Thus, it makes the process even tougher. As these sensors are already expensive and require maintenance to get the correct value, the rural farmers of the villages cannot monitor them properly. That is the constraint of their work.

Moreover, another study [21] highlighted the prospect of integrating AI and UAV in livestock to enhance monitoring. They suggested the use of drone technology with deep learning models to get the best outcome. However, this integration has a high setup cost, which can be expensive for small farms. Additionally, this level of technology needs to be maintained properly with experts who may be unavailable in remote areas. Therefore, there is a constant urge to maintain this technology under supervision to facilitate a better outcome. Furthermore, the paper [13] stated the usage of ICT tools for monitoring livestock. They experimented with the use of modern devices and tried to figure out if this would be a feasible solution for monitoring livestock. The research showed a higher number of people using mobile phones. Thus, it gives a prospect of the study, whereas it also limits the fact that livestock monitoring requires sensors to check environmental factors, which is difficult through only a mobile phone. Also, it is required to check over a huge area to monitor livestock, and that can be problematic when using a mobile phone. Consequently, this can result in adaptation problems that can affect constant monitoring. Therefore, the earlier works focused on IoT-based sensors and integration with AI and ICT tools, which are costly and have constraints on data monitoring.

Overall, many proposed works exist in this monitoring field, but proper execution is a major concern. In our research, we have tried to point out the challenges that these existing proposed methods face and tried to solve them for the proper execution of cattle disease monitoring.

Chapter 3

Dataset

3.1 Data Collection

We have visited many farms and noticed that many animal farms do not record disease information. After realizing the gap, we started our research using secondary data [31]. These include published studies, veterinary records, and trade publications. Our main goal is to analyze the symptoms of diseases and the impact of livestock farms. We have worked with two secondary datasets [31], [36]. One is for extracting more symptoms, and the other is to detract disease. We merged these two datasets and created our custom dataset for better results. With the implementation of our custom data, we can provide insightful analysis and practical advice to improve livestock health management.

3.2 Data Analysis

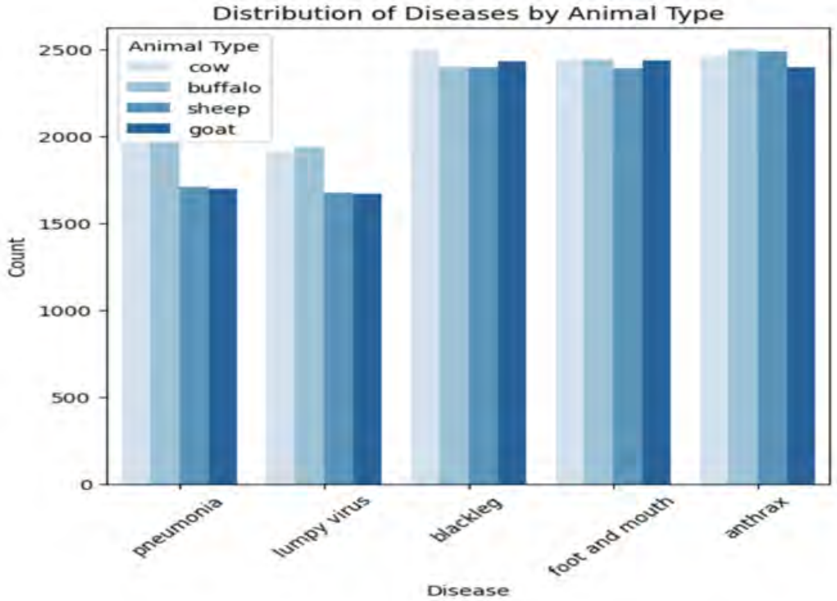


Figure 3.1: Distribution of Disease

The dataset has eight features: animal, age, temperature, temperature status, symptom 1, symptom 2, symptom 3, and symptom 4, and the targeted output is a disease in our updated dataset. Besides, there are 43778 non-null values in the data. There are 28 symptoms to detect diseases. By splitting the dataset into training and testing, we have used 70% of the dataset for training with various models, and the rest of the 30% was used to test the data. In the bar graph of the distribution of diseases by animal type, we see five columns with diseases such as pneumonia, lumpy virus, blackleg, foot and mouth, and anthrax. The count on the vertical axis is 2500. This bar chart represents cows, buffaloes, sheep, and goats, shading light blue to dark blue. Among all the diseases, we can see that blackleg has the highest number of occurrences in cows.

Moreover, anthrax is the most common in buffaloes. On the other hand, the dataset that we used in pre-thesis 2 consists of 7330 data points for pneumonia, 9713 data points for blackleg, 9701 data points for foot and mouth disease, 9842 data points for anthrax, and 7192 data points for LMV virus. Similarly, the dataset comprises six features, with the target output being disease. There is a difference with the updated dataset, as we selected the threshold for temperature there and marked it as temperature status. Additionally, there are four animals with five diseases with 28 symptoms. Therefore, we can easily observe the difference between the updated and original datasets.

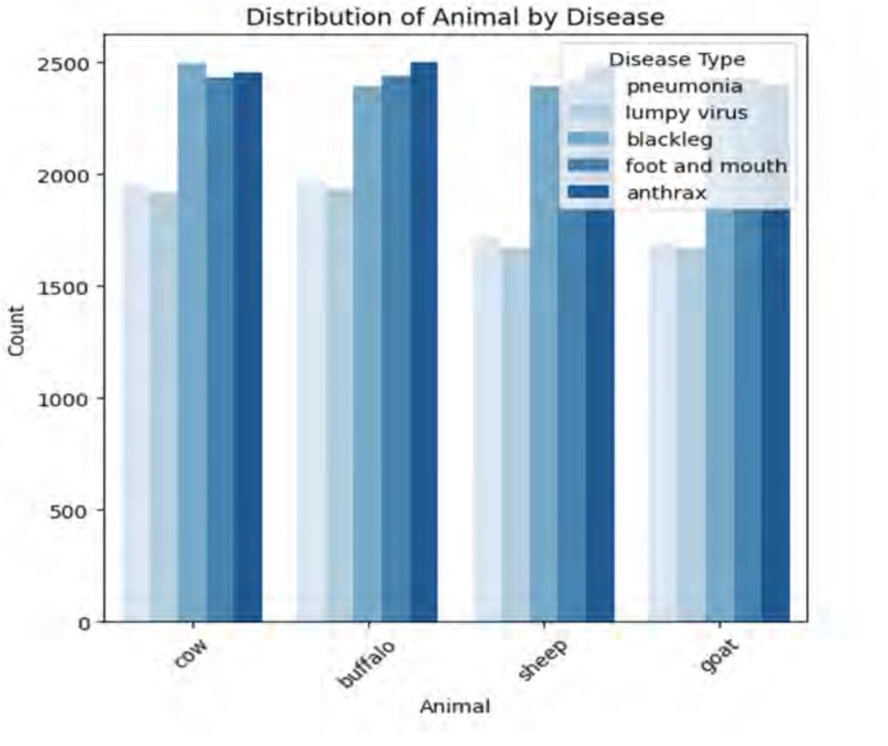


Figure 3.2: Distribution of Animal

Now, if we look closer at the bar graph for the Distribution of Animals, we can observe four columns, each representing an animal species, such as cow, buffalo, sheep, and goat. On the other hand, the count is 2500, as in the previous bar graph. The graph shows that cows are mainly affected by blackleg. Buffalo and sheep are mostly affected by anthrax, and goats are primarily affected by foot and mouth viruses.

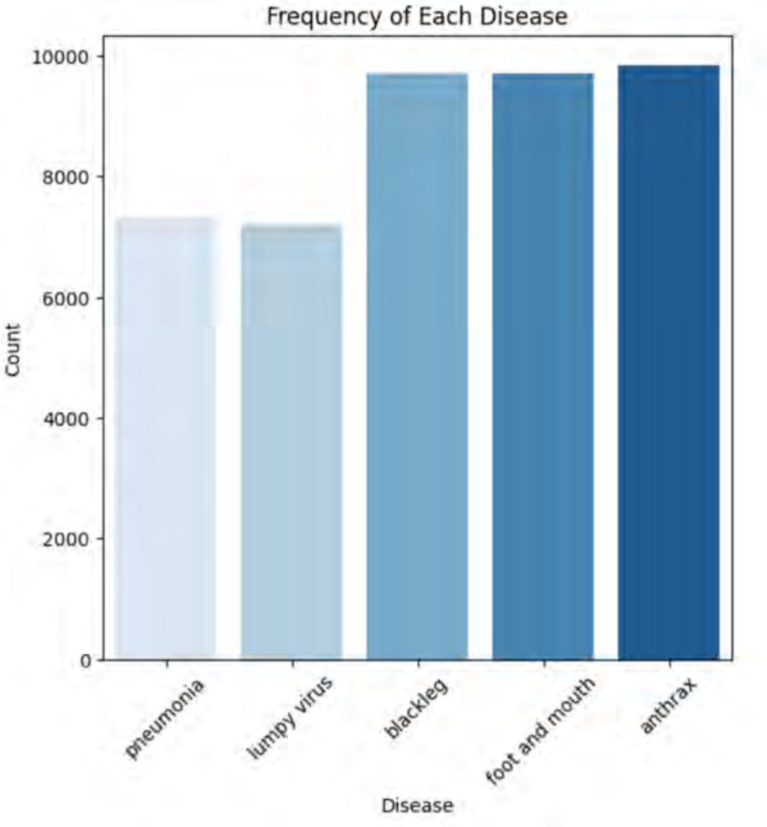


Figure 3.3: Frequency of Disease

This graph showcases the occurrences of each disease when the count is 10,000 in the dataset. Here, we can see that each disease is represented in each column, and anthrax is the most common disease among the five. The incidences of blackleg, foot, and mouth are the same. Besides, the lumpy virus has the lowest number of occurrences. Overall, the frequency of disease is approximately the same in ratio. This indicates that the data classes are balanced.

3.2.1 Data Features

The dataset consists of 8 distinctive essential features, each contributing to an inclusive analysis of diverse aspects of cattle disease detection. For diagnosing disease among cattle (cow, buffalo, sheep & goat), we have taken some crucial attributes like an animal, age, four different types of symptoms, temperature, and temperature status. These features will help to identify cattle diseases (blackleg, foot and mouth, lumpy virus, anthrax, and pneumonia) more precisely. Determining cattle's disease provides straight information about their health and necessary interventions. Animal age is important because their immune system differs at different age levels. The age span of animals in the dataset is 1-15. Also, in every type of cattle disease, their body temperature matters as it is one of the crucial indicators of health status. Inside the dataset, the body temperature range is between 100.1°F to 105°F. From this temperature attribute, the temperature status is driven from, where 100.1°F to 102.4°F indicates normal temperature and from 102.5°F to 105°F means high-temperature status. Moreover, previously, the dataset had three symptom fields, including 24 distinguished health signs. Later, we merged another dataset to bring the 4Th symptom (including painful mouth, dark urine, immunosuppression, blood in urine, etc.) for more accuracy in detecting disease. So, in total, the dataset has around 28 different types of symptoms to point out the disease. Among those, any combination of four characteristics and other features assists in finding the name of cattle disease.

Here are the symptoms that we have merged from the datasets to detect disease accurately:

Symptom	**Description**
Blisters on gums	Vesicles form inside the mouth, on adult females' tongues, lips, and teats, causing feverishness.
Blisters on hooves	High Fever followed by mouth and feet blisters spreading rapidly in herds.
Blisters on the tongue	Fever, and high-rise sores, more so on the top than below the hooves, causing lameness.
Chest discomfort	is caused by actinobacillus lignieresii invading through wounds or minor trauma. Right-sided heart failure leads to peripheral Edema and venous distension, while left-sided heart failure causes respiratory symptoms.
Chills	Cold stress result in muscle shivering, increased heart rate, deeper breathing, and higher body heat when energy stores are low.

Symptom	**Description**
Depression	Occurs when survival instincts are repressed or overridden, indicating genetic variation in resilience to stress factors.
Difficulty walking	Trauma metabolic affliction or diseases like mastitis can cause cows to go down and not stand up. The condition is life-threatening if untreated.
Lameness	Painful condition affecting the locomotor system due to hoof lesions, joint diseases, injuries, or neurologic issues.
Fatigue	Normal consequence of prolonged or intense exercise, acting as a safety mechanism to prevent structural damage.
Loss of appetite	Stress, infection, or disease can cause cattle to stop eating.
Painless lumps	Lumpy skin disease caused by a viral infection spread by biting insects, affecting cattle and water buffalo.
Shortness of breath	Acute bovine pulmonary emphysema from ingesting large amounts of L-tryptophan in lush pasture, causing severe dyspnea and foam around the muzzle.
Sores on gums	are often caused by parasites like mites and lice or viral, bacterial, or fungal infections.
Sores on hooves	Sole ulcers commonly affect beef and dairy cattle, causing lameness and weight-bearing issues.
Sores on the mouth	Fever and blister-like sores on the tongue, lips, teats, and between the hooves, causing severe production losses.
Sores on the tongue	are caused by actinobacillus lignieresii, invading through wounds or minor trauma.
Sweats	Tick-borne toxicosis characterized by Fever, moist dermatitis, and hyperemia, affecting cattle, especially young calves.
Swelling in the abdomen	is caused by gas trapped in the rumen due to natural foaming agents in legumes and grasses, leading to visible swelling on the left side of the body.
Swelling in extremities	Often due to injury or secondary bacterial infection, is common during export voyages.

Symptom	**Description**
Swelling in limb	Edema from excess fluid in tissues, often due to congestive heart failure or venous insufficiency.
Painful mouth sores	Bovine papular stomatitis caused by a parapoxvirus spread through mucous abrasions.
Painful swelling	in the neck is often due to hypoalbuminemia from diseases like Johne's but can have other causes of localized swelling.
Dark urine	Hemoglobinuria with dark brown or red urine from marked hemolytic anemia, leading to a drop in milk production and lethargy.
Blood in urine	Leptospirosis can cause kidney damage and blood in the urine, contracted from contaminated water or sick animals' urine.
Immunosuppression	Decreased immune response to foreign antigens, caused by infectious or non-infectious factors.

Table 3.1: Symptoms and their Description

Feature No	Feature Name	Data Type	Description of Features
1	Animal No	String	Unique identifier for each animal
2	Age	Numerical	Range from 1-15
3	Temperature Status	String	High or Normal
4	Symptom 1	String	One of 28 different symptoms of diseases
5	Symptom 2	String	One of 28 different symptoms of diseases
6	Symptom 3	String	One of 28 different symptoms of diseases
7	Symptom 4	String	One of 28 different symptoms of diseases
8	Disease (Target)	String	Output class

Table 3.2: Data Features

3.2.2 Data Correlation

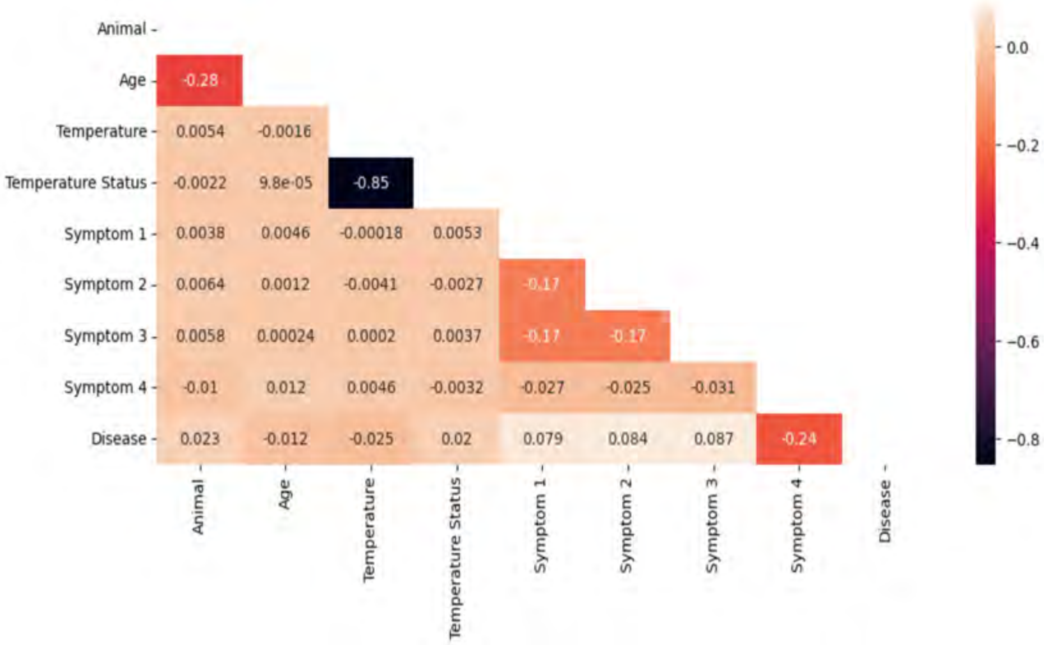


Figure 3.4: Correlation of Custom Dataset

The heat map displayed the relationship between various variables related to animal health. It includes 8 features: animal, age, temperature, temperature status, symptom 1, symptom 2, symptom 3 and symptom 4. Here, in the heatmap, it is important to understand color coding more thoroughly. Because color indicates the power of the relationship among those. The representation of white means no correlation. Lighter colors represent weak relationships among them. While darker colors indicate stronger relationships among them. Relationships among the features can be understood concerning positive correlation and negative correlation. Disease is mostly positively correlated with other features in this heatmap and shown according to the color mapping.

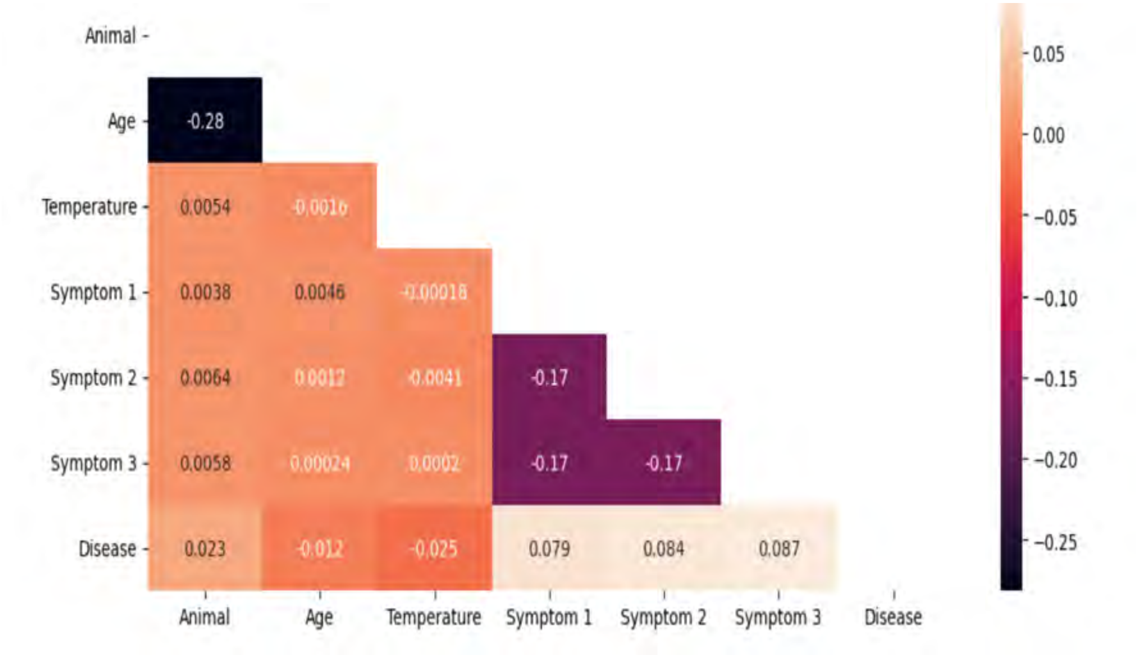


Figure 3.5: Correlation of Default Dataset

Our previous experiment's heatmap displayed the relationship among multiple features related to animal health. It included six features: animal, age, temperature, symptom 1, symptom 2, and symptom 3. Here, the color also indicates the different levels of correlations between those given variables. The heatmap below shows that the relationship between age and animals is strongly negative, according to color. Also, the correlation between disease and symptom 3 indicates a weak positive connection. Other cells represent the relationship between variables according to color codes.

If we compare the heatmap of our previous experiment to the new one, we can see many changes have come to the level of relationship among variables. In the past, when we had a strong negative correlation between age and animals, the new heatmap shows a strong positive connection between them. Moreover, previously, we found a weak positive relationship between disease and symptom 3, and now we are getting weak positive relationships between them. Consequently, we are getting significant differences in heatmap when we change our features to detect cattle health more precisely.

3.3 Data Pre-Processing

The pre-processing starts with label encoding of the data. In the updated dataset, we had categorical data, as all the data were in string format. Thus, we needed a label encoder to convert this into a numerical value. The label encoder converts categorical data into numerical values. In addition to this, we used one hot encoding for our output class. Models like neural networks use loss functions like categorical cross-entropy that actually require the output class to be one hot encoded. It also allows the models to deal with multi-class classification problems.

Moreover, in the datasets, there were no imbalances. As we already had the features along with the targeted value and split the dataset into 70:30, we get the pre-processed data, as the training set contains 30644 data along with the test set with 13134 data. On the contrary, the dataset we used in pre-thesis 2 contains a similar amount of data as the updated one.

We combined the second dataset and added two new columns: Temperature Status and Symptoms 4. We corrected the entries in the Symptoms 4 column using the data from the Temperature Condition column. This process has helped improve our dataset with more accurate and relevant information, enabling more precise analysis and better decision-making about livestock health conditions. This development helps identify correlations between temperature changes and specific symptoms, providing deeper insight into livestock health conditions. This enables more detailed monitoring and the early detection of potential health problems. In addition, this custom dataset will facilitate more comprehensive statistical analysis, making us more effective in identifying trends and anomalies. Altogether, these developments will support the development of better health management strategies and contribute to the overall welfare of livestock.

Chapter 4

Workflow of Methodology

Our research topic is a comprehensive and research-oriented topic that combines machine learning and cattle disease. This topic aims to investigate the symptoms along with factors like temperature and age to provide a disease prediction.

1. Data Collection: Primarily, we wanted to have our primary data, but it was difficult to find disease information of cattle in countries like Bangladesh. Then, we initially worked with two datasets from [31], [36].

2. Data Analysis: The first dataset [31] has instances of nearly 43800 with a total of 6 features. Features are Animal, Age, Temperature, Symptom 1, Symptom 2 and Symptom 3. The disease is the targeted feature. After working with this dataset, we found that with these features, machine learning models could not predict correctly because some output classes were creating difficulty in identifying since they have similar symptoms. We also used correlation to learn about the connections between features. Then, we used another dataset [36] to extract a few more symptoms and customized our previous dataset with two more features. This time, we saw an improvement in our overall output.

3. Preprocessing: First of all, the dataset had no null values. Besides, we have seen a balance between disease and animal data in data visualization, so our dataset has no imbalance. We must use label encoding since the dataset has numeric and categorical data. This is how we process our dataset before applying.

4. Train Test Split: We have a single customized dataset. As a result, we had to split this single dataset into two segments with a ratio of 70:30. Nearly 30,000 data were used to train the models, while around 13,000 data were used to test the model's accuracy.

5. Machine Learning Model Training: We have two segments of model training. At first, we tried to fit our training data to the machine learning models with default parameters and then check the result. Then, we customized the model parameters to get the best result. After getting results from the basic models, we shifted to segment two, the Ensemble Model part. There, we have used voting and stacking methods to gain better results. Since our dataset does not carry a lot of variations, two of our basic models, Neural Network and Gradient Boosting, consistently give the best results. As a result, when we are using the Ensemble model, only around 2% improvement is visible.

6. Result Analysis: To analyze the output, we have multiple segments. At first, we applied some machine learning models to the raw dataset. There, we gained the highest accuracy, which was 83%, and a problem was identified that some classes are difficult to identify. Then, we extract some new features and create a custom dataset in which we see that the result of identifying a disease is higher than the previous one. We used performance metrics like accuracy, precision, recall, and f1 score to compare our models and identify which model provided the best outcome.

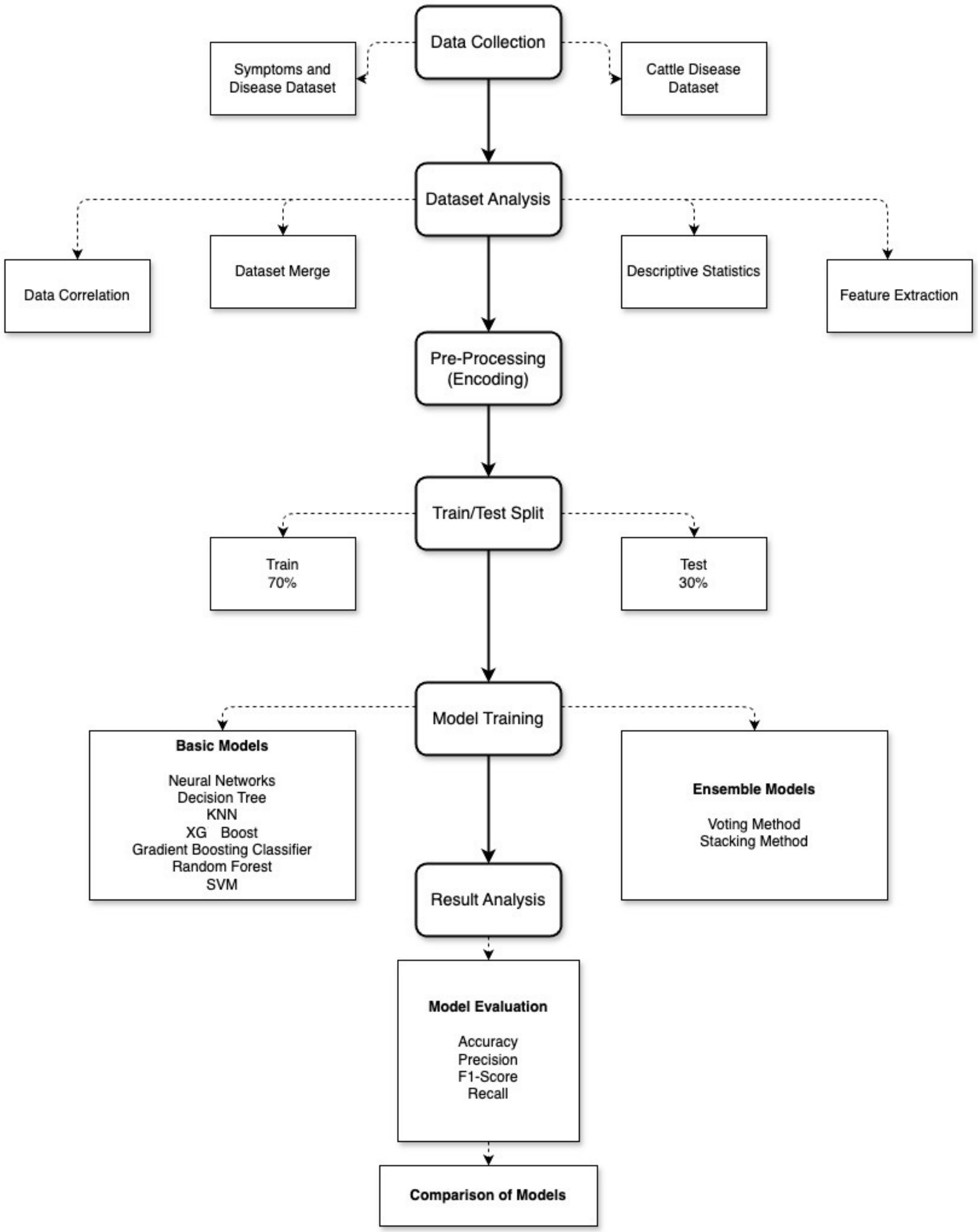


Figure 4.1: Workflow of Methodology

Chapter 5

Description of Utilized Models

5.1 Decision Tree

The decision tree is a strong supervised learning technique for regression and classification tasks. The structure of it is similar to a flowchart where each leaf node holds a class label node. Each internal node represents a test on a property, and each branch represents a consequence. Beginning at the root node, the algorithm uses an Attribute Selection Measure (ASM) to determine the best attribute. It then partitions the set into subsets, generates the decision tree node, and recursively builds more decision trees using the subsets. A leaf node is the last node in the chain. The terms root node, child node, variance, impurity, information gain, and pruning are frequently used in discussions of decision trees. Compared to other algorithms, the decision tree requires less data cleansing and is simpler, more useful for handling decision-related problems, and capable of considering all possible outcomes. It can assist in problem-solving and help make decisions in real life.

There are some steps on how the decision tree works.

1. There will be a root node, which is S. This root node will contain whole datasets.

2. After selecting the dataset, we will use attribute selection measures to identify the best attribute.

3. , we have to separate the root node S into subnodes.

4. After separating the sub-nodes, we must create a decision tree node with the best attribute at its center.

5. After using the dataset, we have to go again to step 3 and recursively construct a new decision tree.

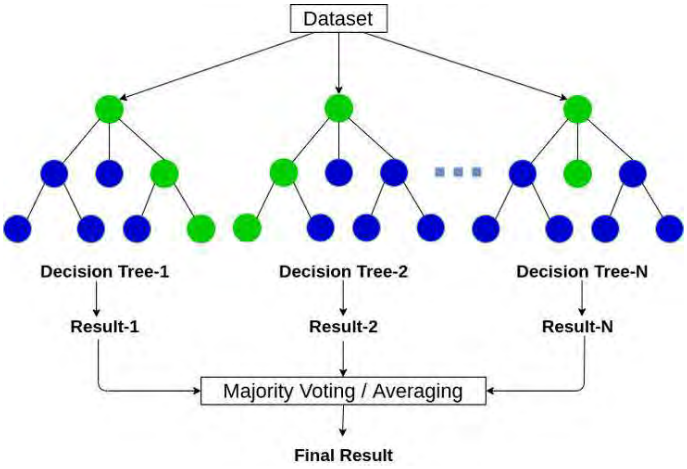


Figure 5.1: Diagram of Decision Tree [34]

5.2 Random Forest

Random forest is a modeling and behavior analysis tool based on decision trees. The most widely supported forecast is the one that receives the most votes, and it makes use of several trees to reflect different instances of data classification. Random forests are helpful for complicated jobs because they can swiftly handle variables, balance big data sets, and handle them efficiently. They also give estimates of the relevance of variables and provide a better way to handle missing data. The random forest algorithm operation is explained in the phases that follow:

Step1: We must select a dataset from a given training dataset and choose random samples.

Step2: This algorithm will build a decision tree based on the training dataset.

Step3: The decision tree will be averaged to determine the main answer.

Step4: The best outcome will be the most voted result.

We refer to this combination of various models as a common goal.

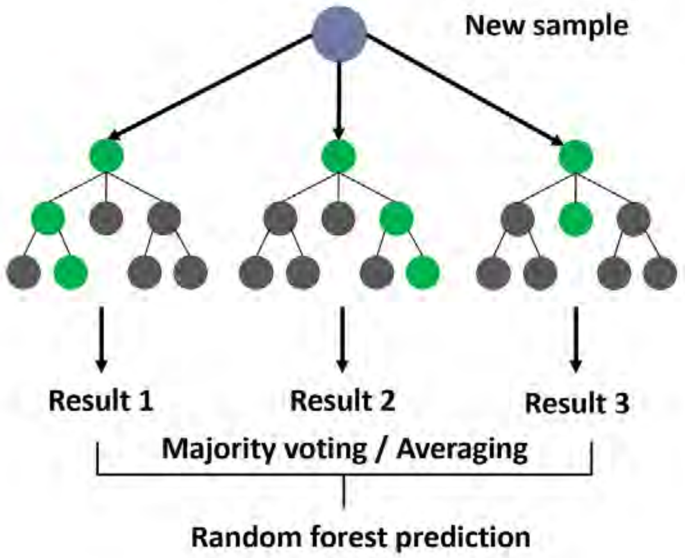


Figure 5.2: Random Forest [34]

5.3 XGBoost

XGBoost is an open-source machine-learning tool that combines decision trees and gradient boosting to improve model accuracy. It is part of the gradient boosting algorithm family, which combines the predictions of multiple weak learners, typically decision trees, to create a stronger model. XGBoost enhances traditional gradient boosting by incorporating optimization and regularisation techniques. It takes in training data, trains a model, and evaluates it on new data until it stops improving. XGBoost uses bagging, training multiple decision trees, and combining their results, making it faster and more effective in situations with multiple features.
Initially, we must create a single tree leaf and use the similarity score to generate a decision tree. Generate the average of the target variable as an estimation and then construct the leftovers using the selected loss function. Homogeneity is improved by picking a suitable node based on the similarity score. Compute the data obtained by breaking the node at a specific location. The information obtained distinguishes the previous and new similarities in the data. Construct a tree for the desired length, then use the regularization hyperparameter to reduce and regularize it. Predict the residual amounts using the Decision Tree, calculate the new set of leftovers using the development rate, and continue the procedure for each tree.

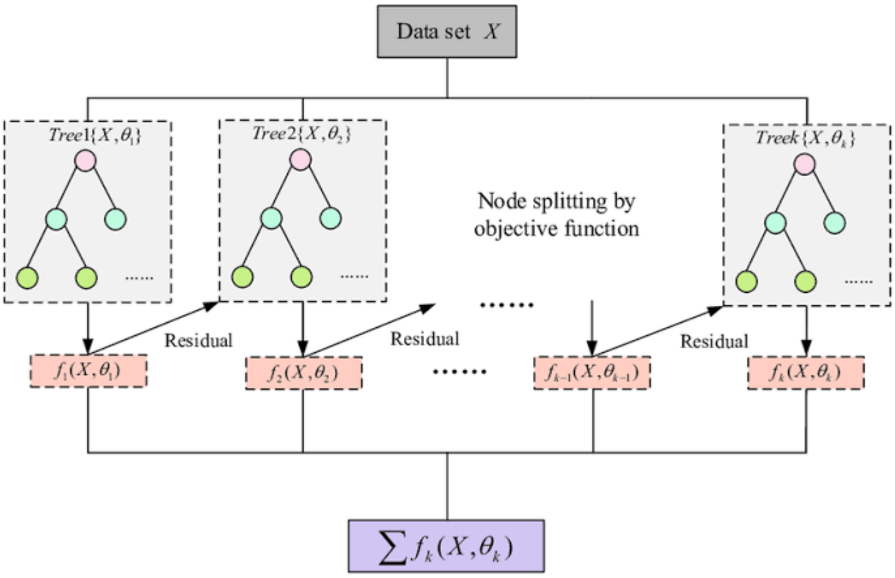


Figure 5.3: Diagram of XGBoost [8]

5.4 K-Nearest Neighbors Algorithm

A monitored, irregularly trained classifier, the K-Nearest Neighbors (KNN), examines the description or value of a fresh data point based on similarities. Using the similarities theory, the distance between each new data point in the test dataset and each data point in the training dataset is determined. Depending on the task and data, the method uses an array of distance indicators, including Manhattan, Euclidean, and Minkowski distances.

When any part of the K nearest neighbors has been established, the algorithm makes predictions using its descriptions or associated values. The predicted result in the case of regression is the averaged or weighted mean of the results of the K neighbors; in classification problems, the expected label is the majority class among the K neighbors.

The KNN method is simple to set up due to its low complexity. As all of the data has been saved in memory, the algorithm can adjust to fresh instances and help with predictions for the future. The only variables needed for learning a KNN algorithm are the value of k and its distance metric selection according to the evaluation metric. This approach tends to be used in classification problems when similar points can be found in close range.

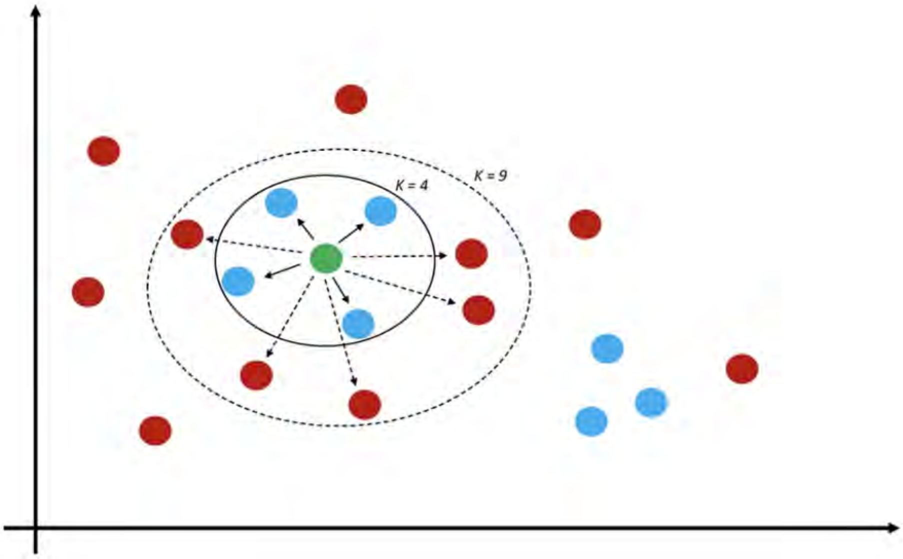


Figure 5.4: Diagram of KNN [10]

5.5 Support Vector Machine

SVM (Support Vector Machine) is a supervised machine learning algorithm for classification and regression tasks. SVMs work by identifying an optimal hyperplane that separates distinct groups in a high-dimensional feature the distance between the nearest data points of each group and the hyperplane because the SVM algorithm aims to enhance the margin accuracy. In situations like complex decision thresholds, SVM methodology is remarkably efficient. Moreover, it can handle linear bonds by employing kernel, and non-nonlinear is highly known and used for its ability to adapt and generalize effectively. They are also strong and versatile in various aspects, such as image recognition, bio-informatics, text classification, email classification, and gene classification. The SVM model is impressive for dealing with vast amounts of data. Still, at the same time, it shows a lower tendency to recall information too much, a prevalent characteristic of a machine learning model. Above all, SVM is a top-performing model in correlative analysis that highlights its effectiveness and reliability in various machine learning applications.

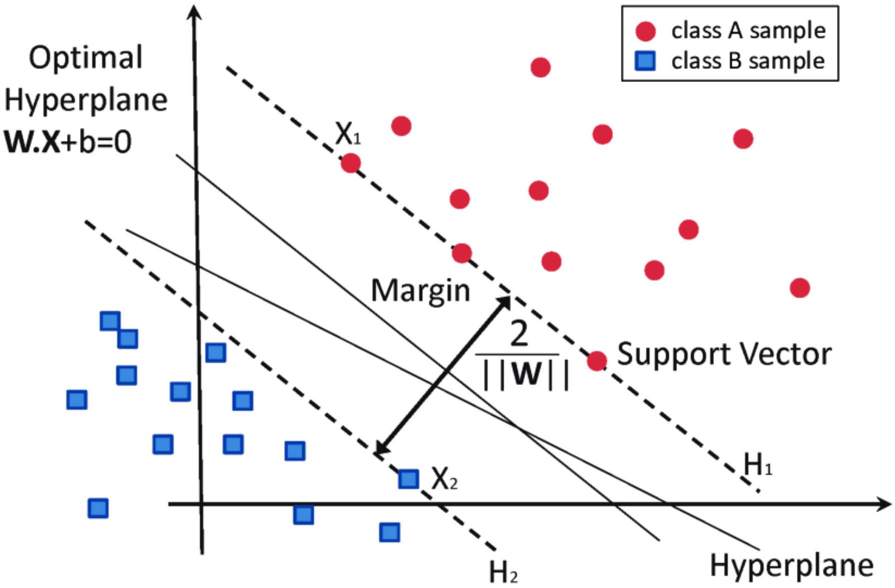


Figure 5.5: Diagram of Support Vector Machine [3]

5.6 Gradient Boosting Classifier

One of the most advanced and powerful machine learning algorithms is known as GBC (Gradient Boosting Classifier). It is a part of the broader family of assembled learning techniques mainly used for classification tasks. This technique aims to build up the predictive performance by combining the power of multiple base learners. Gradient Boosting utilizes a sequence of weak learners, such as decision trees, that make better accuracy, optimizing the faults of preceding models iteratively. This process begins by configuring the model with a simple prediction that is either the mean of the target variable for the regression task or a single class probability for the classification task. Improvements can be made from the baseline provided by these initial models. In each step, the residual errors of the combined ensemble of all previously trained trees are predicted by training a new decision tree. These residuals represent the gap between the predicted and observed values. It indicates the area where the present model is underachieving, minimizing the loss function, which is another aim of GBC. To reduce the prediction errors, the new tree is aligned with the anti-gradient of the loss function. Before adding to the model, predictions are scaled by a learning rate. We get more accurate and precise results from the smaller learning rate, but it needs more iterations. We add trees in this process until a fixed number of iterations or when there is minimal improvement through step-by-step correcting errors in the model. Key parameters (Number of trees, learning rate, max depth of trees, subsampling) in the GBC help us get our desired output.

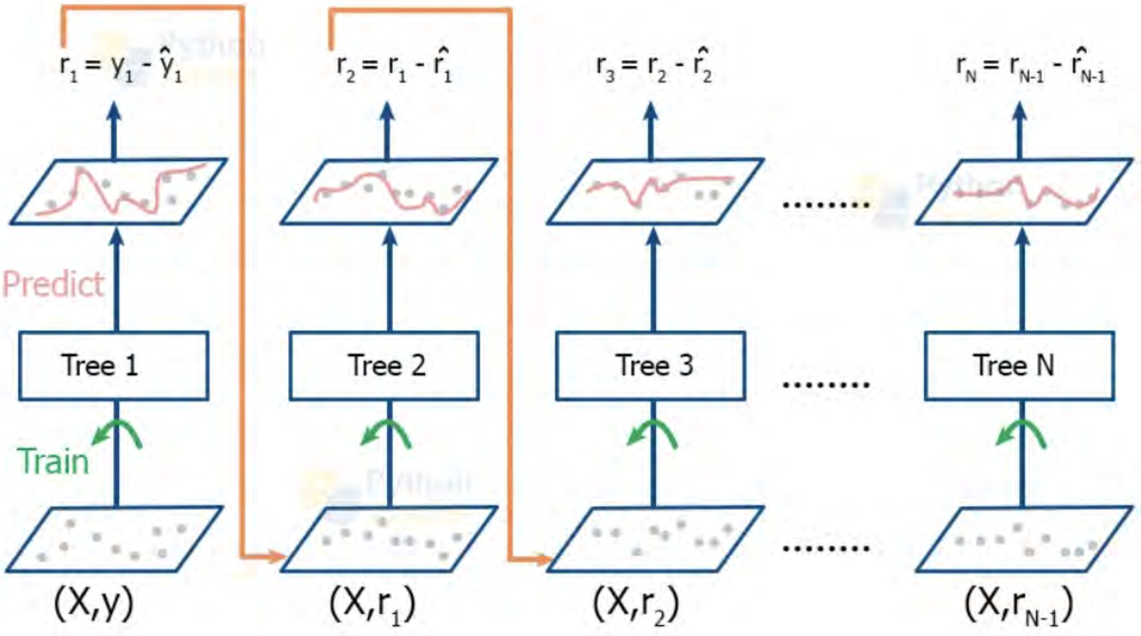


Figure 5.6: Diagram of Gradient Boosting Algorithm [15]

5.7 Neural Network

Neural networks (NNs) are a specific forecasting model used for data processing and forecasting. Comparable to response surface methods, an embedded neural network maps input parameters to a particular output and produces more consistent results than traditional mathematical analysis methods, such as regression analysis, while requiring significantly less computer work. NNs work similarly to a network of biological neurons. The synthetic nerve cell is considered the building block of NN and a quantitative model that mimics the behavior of a biological neuron. The input is passed through the synthetic nerve cell, and the output is generated after non-linear processing. Additionally, weights are added to the input parameters before they reach the neurons to simulate the casual character of biological nerve cells [25]. Three primary processes are required to construct a neural network: first is to build the structure of the NN. The second is to define the preparation method necessary for the development phase of ANN, and the third is to identify the mathematical functions that describe the quantitative model. The training phase of the neural network plays a significant role in decision-making as it selects the best weight and minimizes the loss function. NN empowers computers to make intelligent decisions with minimal human intervention. This capacity enables us to tackle tasks beyond the capabilities of conventional programming. As a result, different NNs adopt different development methods. Overall, this model gives us the best result.

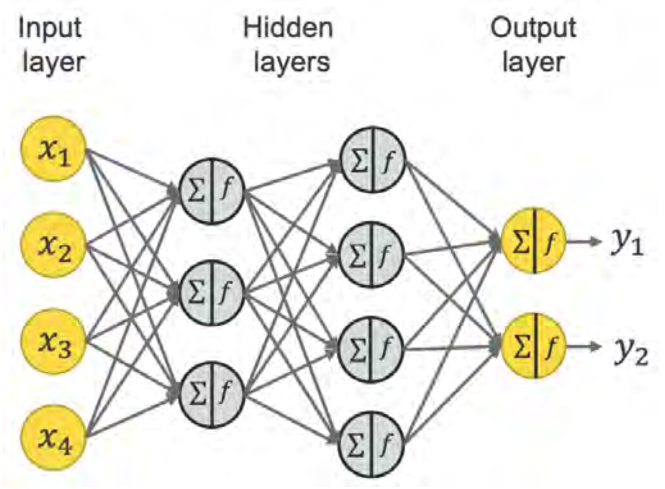


Figure 5.7: Architecture of Neural Networks [35]

5.8 Ensemble Model

To detect cattle disease, we have executed several machine learning models two times; one experiment before our customized dataset and another one after our customized dataset. In both tests, we have run seven models, which are the neural network, Knn, gradient boosting classifier, XGBoost, decision tree, random forest, and SVM, to identify which process can do the disease detection more accurately and precisely. Previously, we had customized our dataset with 6 features to detect cattle disease. After running those above-mentioned machine learning models, we compared the ability to detect cattle disease for each model within themselves. We saw that among those 7 models, the Neural Network has the highest accuracy rate (82.87%), precision (82.85%), recall (82.87%), and F-1 score (81.18%) than other models. However, the methodology SVM has shown the lowest rate of accuracy (70.68%) and others. So, we can say that, before customizing our dataset, Neural Network is the best machine learning model for detecting cattle disease from our given features, and on the other hand, SVM is the worst of all models. Although we got a good accuracy rate in our previous dataset, the reliability rate was not promising enough for our goal. So, we have customized our dataset and added two more features for a more precise result. After running those seven machine learning models in our customized dataset, we have analyzed the accuracy and other rates, which have greatly improved. Here, we have seen that the Neural Network has performed better in detecting cattle diseases than other models. It has an accuracy rate of 89.45%, which is 6.58% more than our previous experiment. Thus, the Neural Network is best again in detecting cattle diseases from our updated dataset. Also, we can optimize that SVM is not giving us a promising result, and at the same time, the accuracy rate has lessened from the result of the non-customized dataset. So, we can say that the SVM model is still the worst one.

We have worked through the ensemble models, such as the voting and stacking methods. The voting method combines predictions from multiple models to make a final decision. It either uses majority voting for classification or averaging for regression to improve the whole predictive performance and decrease errors. On the other hand, the stacking method includes training several base models and then working on their predictions as input features for a meta-model. Then, this Meta-model learns to integrate those predictions to enhance the aggregate predictive performance. Again, the voting method has two types; one is soft, and the other is hard. We used the soft voting method and got an overall accuracy of 91%.

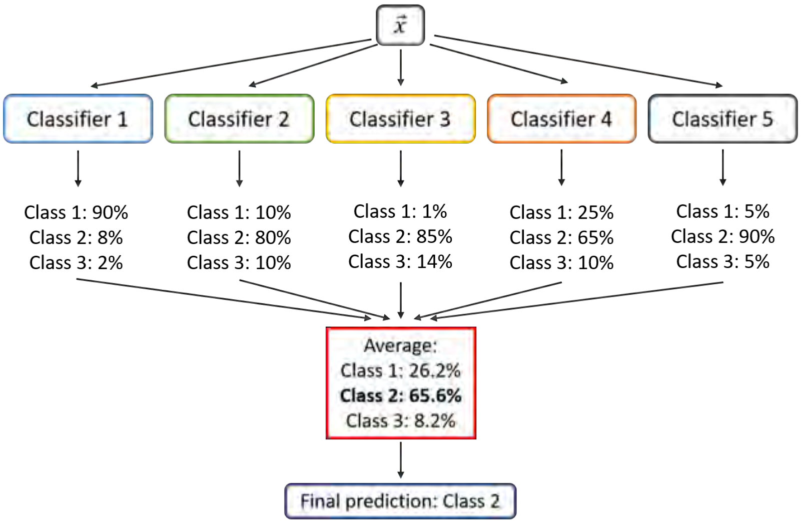


Figure 5.8: Diagram of voting method [28]

However, using the stacking method, we have 92% output as an accuracy rate. Previously, we have gained 89% of accuracy by using Neural Network. Since our dataset has fewer variations between all the data we are working with such data where some models are consistently giving the best result, so the ensemble model has not provided a huge improvement. It has improved by about 2% from our previously gained models. Our primary motivation was to create an ensemble model that could combine the strengths of individual models and mitigate the weaknesses of those models individually. In the future when we implement this model for end users through mobile applications, this ensemble model can help us as it has the highest accuracy in detecting disease.

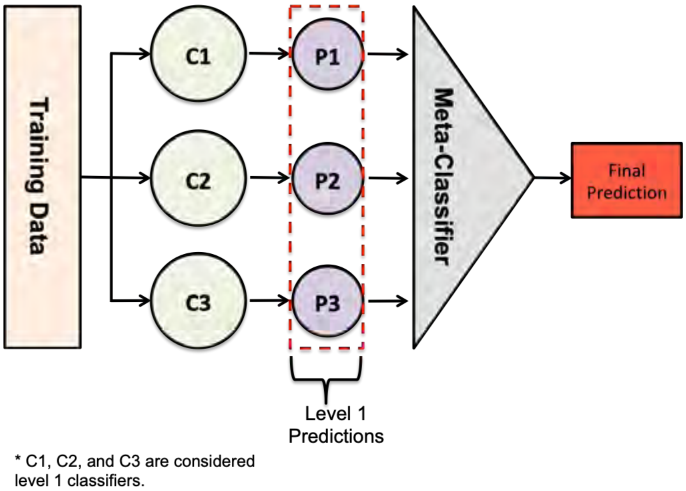


Figure 5.9: Diagram of stacking method [7]

Chapter 6

Experiments and Findings

6.1 Experimental Settings

We have chosen to work in the Google Colab platform, a cloud-based environment. We have stored all our data in Google Drive. Our dataset is a CSV file. For computational tasks, we have used the free access of GPU shared by Google Colab, and since we worked with text data, we do not have to face difficulties regarding the environment. We used pandas for data visualization, analysis, and manipulation. Besides, we used numpy for computational benefits. Sk learn was used for data processing and several machine learning algorithms. Tensorflow, along with Keras, was used for model development and training. Additionally, we have to use feature tuning to gain better output. Overall, in this way, we have set the environment for our research work.

6.2 Evaluation Metrics

Confusion Matrix

It is used to determine the overall performance of the model or classifier. The matrix is a NxN matrix that provides the summary of actual and predicted values. By looking at the matrix representation, we can see how many instances are predicted correctly or how many errors there are. In our research work, we have seen that neural networks and gradient boosting are the models that provide the best results in both scenarios, and SVM is the model that provides the worst results. Overall, the confusion matrix gives us an idea of actual data's correct and incorrect predictions.

Accuracy

It helps determine a model's performance. It represents the ratio of correct predictions to the total number of predictions. Precision: Precision is another performance indicator of a model that shows the percentage of correct predictions generated by the models.

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + False\ negative + True\ negative + False\ positive} \tag{6.1}$$

Precision
Precision is widely used as a machine learning classification model performance metric. It indicates the optimistic predictions that the model measures. Generally, it means the percentage of true positives predicted by the model out of all the positive(true positive and false positive) predictions made by the model. However, the 'weighted' term is used for a multi-class classification. As a result, the weighted average is determined using the number of actual rows for each label, and after this, each label's precision is identified. Therefore, precision helps us identify if the model's accuracy is good enough by checking the optimistic predictions.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \tag{6.2}$$

Recall
A model's recall is its capacity to forecast a positive outcome reliably. It is the proportion of all positive predictions to positive projected outcomes. 'Weighted' refers to average in the context of multi-class classification. It is determined by considering each class's support while calculating each label's average.

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \tag{6.3}$$

F1 Score
The F1 score is calculated using the harmonic mean of recall and precision. Model recall and precision scores are considered when calculating the value, which is determined by the harmonic mean and aids in model comparison.

$$F1 = 2.\frac{Presicion.Recall}{Presicion + Recall} \tag{6.4}$$

6.3 Results on Default Dataset

Model Name	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Neural Network	82.87	82.85	82.87	82.18
KNN	79.10	80.13	79.10	79.59
GBM	81.90	81.88	81.90	81.85
XGBoost	78.11	78.10	78.11	78.11
Decision Tree	77.58	79.27	77.58	77.42
Random Forest	79.67	81.30	79.67	77.27
SVM	70.68	73.68	70.68	70.66

Table 6.1: Results on Default Dataset

Neural Network With the Neural Network model, our model correctly classified 82.87% of the instances. Our precision is 82.85%, indicating that of all the instances we predicted as positive, 82.85% were positive. Our recall stands at 82.87%. The F1 score, balancing both precision and recall, is 82.18%.

KNN (K-Nearest Neighbors) Using the KNN model, we attained an accuracy of 79.10%, which means our model correctly classified 80.13% of the instances. Our precision is 79.10%, and our recall is 79.10%. The F1 score for this model is 79.59%.

Gradient Boosting Classifier For the Gradient Boosting Classifier, we reached an accuracy of 81.90%. Our precision is 81.88%. Our recall matches our accuracy at 81.90%, indicating that we correctly identified 81.90% of the actual positive instances. The F1 score is 81.85%.

XGBoost With the XGBoost model, we got an accuracy score of 78.11%, indicating that our model correctly classified 78.11% of the instances. Our precision is 78.10%, and our recall is 78.11%. The F1 score for this model is 78.11%.

Decision Tree Using the Decision Tree model, we achieved an accuracy of 77.58%. Our precision is 79.27%, and our recall is 77.58%. The F1 score for this model is 77.42%.

Random Forest For the Random Forest model, we attained an accuracy of 79.67%. Our precision is 81.30%, meaning 81.30% of the instances we predicted as positive were indeed positive. Our recall stands at 79.67%, and the F1 score is 77.27%.

SVM (Support Vector Machine) With the SVM model, we achieved an accuracy of 70.68%. Our precision is 73.68%, indicating that 73.68% of the instances we predicted as positive were truly positive. Our recall is 70.68%. The F1 score for this model is 70.66%.

Overall, the Neural Network model gives the best accuracy at 82.87%. The SVM model gives the lowest accuracy at 70.68%.

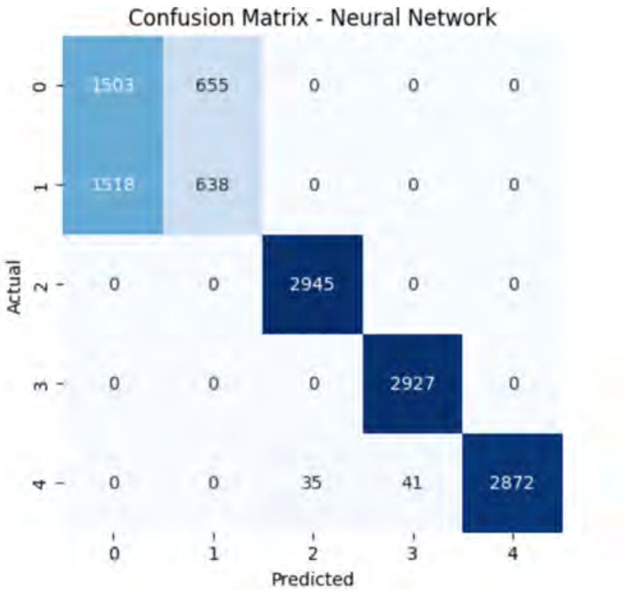


Figure 6.1: Confusion Matrix of Default Dataset for Neural Network

The accurate values are the diagonally presented values, and the rest are the errors. If we look at classes 0 and 1 here, there are many errors as they were not predicted correctly.

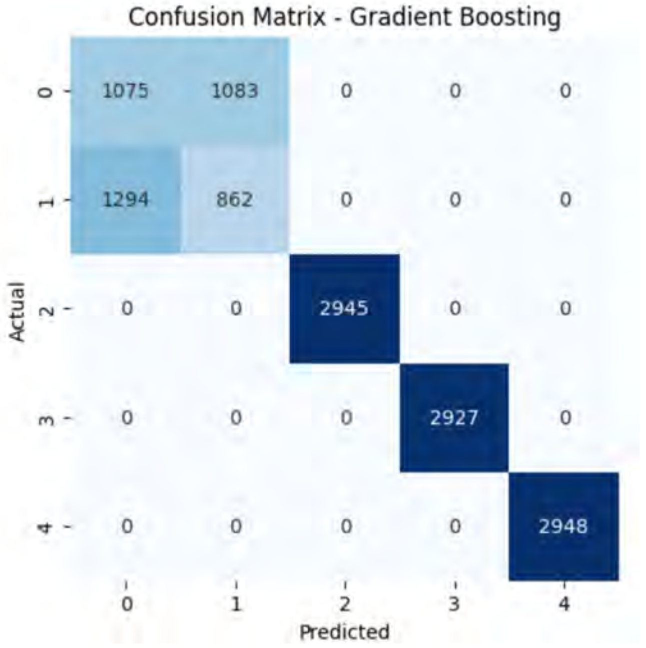


Figure 6.2: Confusion Matrix of Default Dataset for Gradient Boosting

6.4 Results on Custom Dataset

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Neural Network	89.45	89.46	89.45	89.44
Knn	84.81	85.65	84.81	85.19
GBM	87.87	88.15	87.87	88.01
XGBoost	88.26	88.43	88.26	88.59
Decision Tree	83.35	83.40	83.35	83.18
Random Forest	85.29	87.20	85.29	85.53
SVM	73.00	76.04	73.00	73.30

Table 6.2: Results on Custom Dataset

Neural Network With the Neural Network model our model correctly classified 89.45% of the instances. Our precision is 89.46%, indicating that of all the instances we predicted as positive, 89.46%were actually positive. Our recall stands at 89.45%. The F1 score, balancing both precision and recall, is 89.44%.

KNN (K-Nearest Neighbors) Using the KNN model, we attained an accuracy of 84.81%, which means our model correctly classified 84.81% of the instances. Our precision is 85.65%, and our recall is 84.81%. The F1 score for this model is 85.19%.

Gradient Boosting Classifier For the Gradient Boosting Classifier, we reached an accuracy of 87.87%. Our precision is 88.15%. Our recall matches our accuracy at 87.87%, indicating that we correctly identified 87.87% of the actual positive instances. The F1 score is 88.01%.

XGBoost With the XGBoost model, we got an accuracy score of 88.26%, indicating that our model correctly classified 88.26% of the instances. Our precision is 88.43%, and our recall is 88.26%. The F1 score for this model is 88.59%.

Decision Tree Using the Decision Tree model, we achieved an accuracy of 83.35%. Our precision is 83.40%, and our recall is 83.35%. The F1 score for this model is 83.18%.

Random Forest For the Random Forest model, we attained an accuracy of 85.29%. Our precision is 87.20%, meaning 87.20% of the instances we predicted as positive were indeed positive. Our recall stands at 85.29%, and the F1 score is 85.53%.

SVM (Support Vector Machine) With the SVM model, we achieved an accuracy of 73.00%. Our precision is 76.04%, indicating that 73.00% of the instances we predicted as positive were genuinely positive. Our recall is 73.00%. The F1 score for this model is 73.30%.

Overall, the Neural Network model gives the best accuracy at 89.45%. The SVM model gives the lowest accuracy at 73.00%.

The Confusion Matrix below is based on our custom dataset.

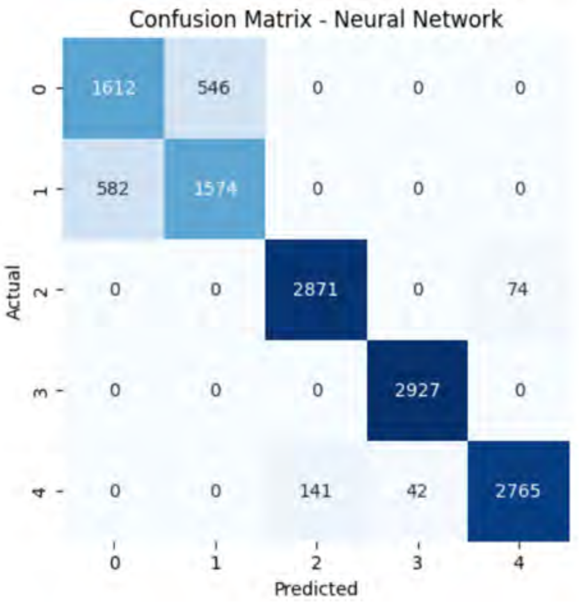


Figure 6.3: Confusion Matrix of Custom Dataset for Neural Network

However, in this case, the neural network has improved. For class 1, it now correctly predicts 1574 instances. Gradient boosting also shows fewer error values, as this model also provides better results.

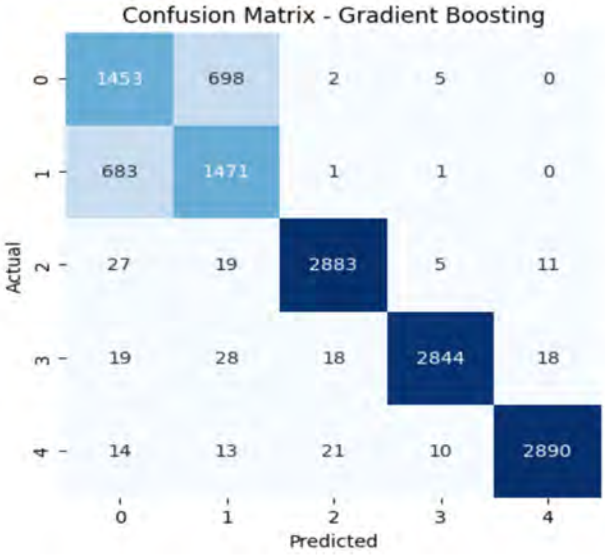


Figure 6.4: Confusion Matrix Custom Dataset for Gradient Boosting

Gradient Boosting also performed better in the custom dataset than our default dataset.

6.5 Comparison and Experimental Findings

We divided our overall experiment into two segments. First of all, in segment 01, we have used our first dataset [31], where we have gained the best output by using a Neural network and gradient boosting classifier. However, we have detected that some output classes cannot be appropriately identified in this dataset due to a lack of features. In particular, lumpy viruses and pneumonia are difficult to detect in the actual class. As in this dataset, there are three fields of symptoms, but in most cases, the symptoms are similar for both diseases. So, the overall performance was moderate this time. After that, we tried to determine how to increase the performance to detect diseases more accurately. This moderate performance was the issue since we worked with only 6 features. Then, from this dataset, [36] have extracted 4 new symptoms of those 5 diseases. We have included a new column, and those symptoms are validated in the dataset. So, this time, we have 8 features to work with. We have also converted the Temperature feature into a new feature named temperature status, where we created two class names, High and Normal. We used this feature to include the value of the symptom 4 table. According to the paper [37], average temperatures of cattle are 100°F-102°F, and anything over it will be considered as sickness. So, the mean value of the temperature class was considered a threshold point based on the 'High' and 'Normal' classes. This time, when we fit our dataset in the machine learning models, we have gained better performance. We have achieved the highest accuracy of 91% by using neural networks and gradient-boosting classifiers. As a result, both of these diseases previously provided the best results. By imposing some extra features, we have gained the best performance. So, if we can add more features in the future, the chances of accurately detecting disease will increase, and it can be essential for livestock farms to mitigate their losses.

We have also tried to implement ensemble models using voting and stacking methods. We have customized the parameters, but we have gained the highest accuracy of 92%. Since our dataset has less variation and models like neural networks and gradient boosting constantly give us the best results, the overall difference between ensemble and basic models is not the same. If we can create more features and variations in our dataset, these ensemble models may provide more improvements.

Overall, the improvement of disease detection is highly dependent on features. If we can primarily collect data with more and more features, it will be beneficial for the machine learning models to find the best pattern for detection. In our future work, we will try to include more diseases to detect them with more and more features.

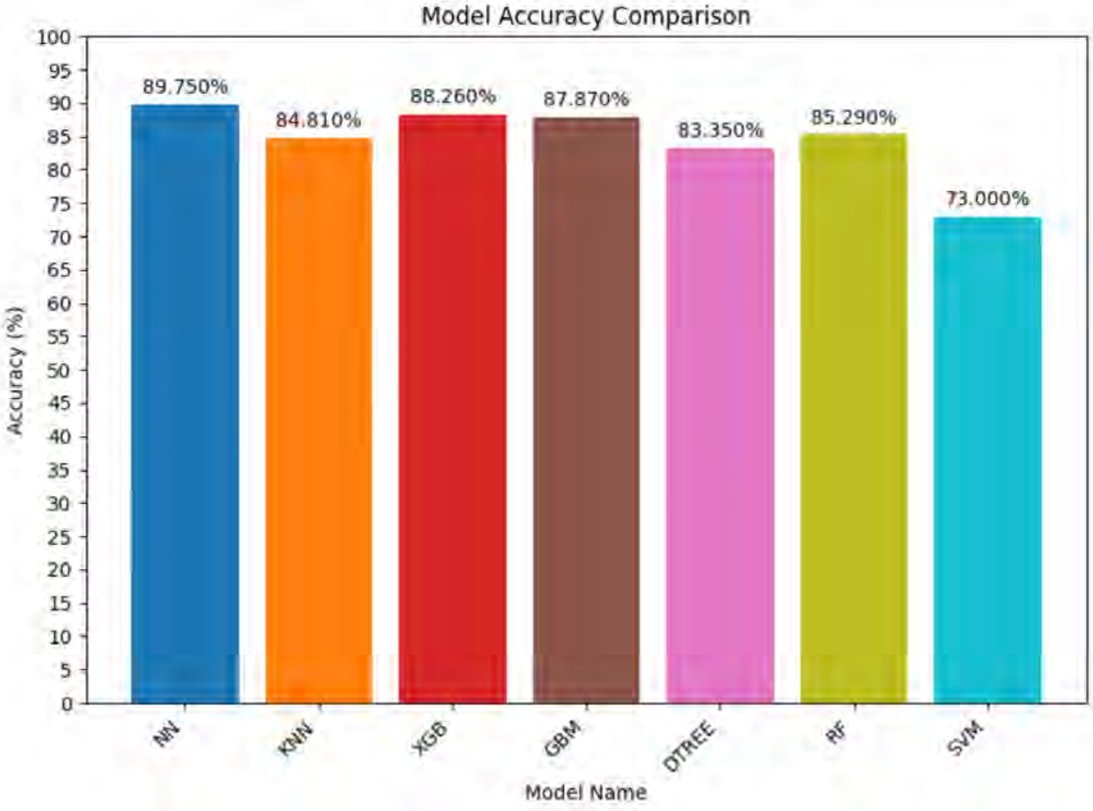


Figure 6.5: Model Accuracy Comparison Diagram

This final bar graph visualizes our models' overall performance. After using the custom dataset, we gained better overall output. This bar graph shows that the Neural Network and Gradient Boosting Classifiers provide better performance. We have tried customizing those models by modifying parameters to get the highest result. Meanwhile, SVM is the model that did not perform well, as shown in the bar graph.

Chapter 7

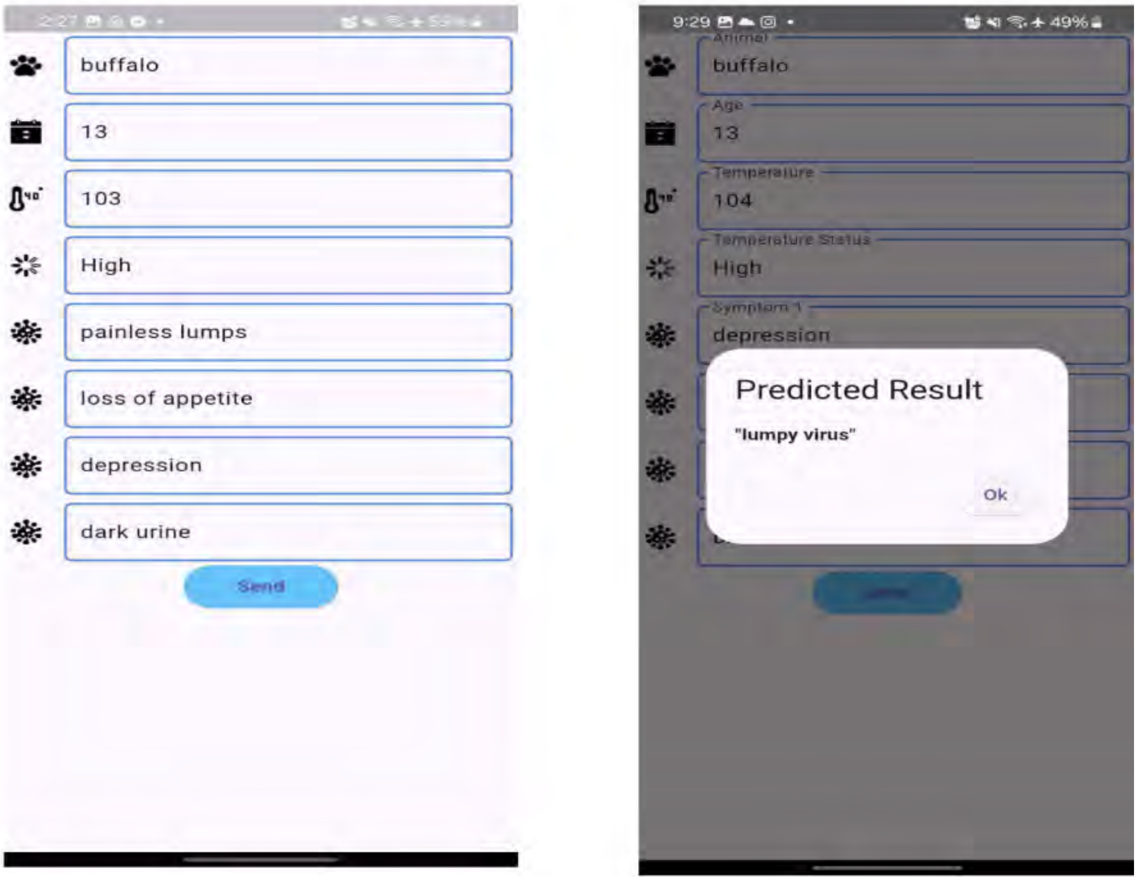
Discussion

7.1 Significance of our research

Our research work is a unique approach to detect and monitor cattle disease. In our research, we have primarily focused on symptoms to detect diseases with higher accuracy. Our primary focus was to make the monitoring process more accessible and reduce the impact of this disease on the livestock industry. Initially, we could not differentiate between two diseases with our dataset in our research work. Then we realized that since we have less number of features to work with, it is not enough for the machine learning models to detect disease accurately. As a result, we extracted some new features and created the custom dataset by adding these features. This time, all the diseases were predicted with better results. This is what happens in real life. Most farms provide medicines by looking at common symptoms, but that does not bring any solution because they sometimes misjudge the disease. It is not possible for farmers or sometimes even doctors to detect disease initially by looking at symptoms because multiple diseases have the same symptoms. Besides, there are almost 200 diseases. So, the overall process is complicated for humans to solve the issue directly. That is why we have tried to solve this problem by using machine learning. If we provide more symptoms or features, our models can accurately learn the disease pattern and give the output. Overall, this is such research work where we have tried to mitigate the impact of cattle disease with the help of machine learning. Besides, it will make the overall monitoring process more accessible.

7.2 Prototype

After implementing machine learning models, we considered implementing our work to the end users. We concluded that implementing our work with a mobile application would be a better solution. As Bangladesh is becoming digital daily, almost every farm has at least one member who uses a smartphone. So, we decided to make an app that the farmers can use to provide immediate input. Our app has the functionality of taking inputs, and after having all the inputs, it generates results based on the machine learning model we already trained. Based on that, the app will show the output directly. It is a prototype of how we will implement our research work in the future.



buffalo

13

103

High

painless lumps

loss of appetite

depression

dark urine

Animal: buffalo

Age: 13

Temperature: 104

Temperature Status: High

Symptom 1: depression

Predicted Result

"lumpy virus"

Figure 7.1: Prototype Design of our Application

7.3 Challenges

Our research work is highly dependent on data. The more data we get we can fit our models better to detect the patterns of each disease. Initially, we failed to get enough data. By extracting features from our custom dataset, we have tried to reduce the problem. Besides, we have faced some difficulties in our application, so we have not finished making the overall application. First, we thought we would make our application based on cloud data to save weight. However, getting results from our app will require a stable internet connection. Besides, there are many farms in rural areas. So getting a stable internet will be difficult. So, we do not approach it that way. We trained our model and fit the model in the backend to get the result. Besides, we are still facing an issue of dynamic IP. We are using NGROK, which provides dynamic IP. And we get a new IP every time we close the session and restart. As a result, our generated app stops working.

In the future, we will try to fix all those issues and make the UI more accessible for users to provide input. For instance, we plan to use a drop-down for the Animal and Temperature Status field to reduce errors while filling in the inputs. Besides every farm needs to have a unique ID and account so that they can log in and see the overall status of their cattle but we have not included this feature yet. In addition to this, we plan to include a medical history feature in our app. However, language can be a barrier since in rural areas in Bangladesh, they are more comfortable with Bangla than English so we have a plan for that as well. Overall, our goal is to create a user-friendly mobile application for end users from rural and urban areas that has stable performance.

Chapter 8

Conclusion and Future Work

Our primary research purpose was to determine diseases so that their impact can be reduced correctly. While collecting data, we observed that only a few livestock farms were aware of these impacts. The traditional method is to identify a disease based on symptoms. Farm laborers know a few symptoms; based on that, they normally try to assume the disease.

In our research, the dataset we used has prevalent symptoms that help veterinary doctors make decisions quickly to reduce the impact. We also wanted to encourage the farms to collect data to identify diseases swiftly and correctly. For example, analyzing multiple symptoms helps the machine-learning model identify diseases with more accuracy.

In the future, we want to focus on many more diseases. Currently, we only detect 5 major cattle diseases. Still, there are almost 200+ diseases that are causing harm to livestock farms, so we want to focus on more diseases and try to identify them correctly. Besides, all these diseases significantly impact the dairy market. For example, cattle disease severely impacts milk, meat, and leather production. Besides all these diseases, they are critical factors in the supply and demand of the livestock market. So, future work will analyze the supply-demand dynamics of the market and try to detect more diseases to help livestock farms gain more profit, fulfill customers' demands, and create stability in the livestock market.

In addition, as our research was mainly focused on how we can detect diseases and reduce their impact, we plan to implement our work to the end users. In Bangladesh, almost every farm owner, sometimes even the farmers, has a smartphone. We utilize this as a medium to interact. We plan to develop a simple mobile application by which farmers or farm owners can input inputs, and based on our models, it will provide the farmers with a possible disease. This application will be user-friendly because we aim to create an application where models will already be trained. As a result, it will not take much time to provide suggestions to the farmers. In Bangladesh, this plan has not yet been implemented. As a result, we hope that, after proper implementation, it can be a massive boost for the livestock farms to handle cattle disease.

Our research has proven its significance in exploring machine learning techniques for livestock disease detection, which showed promising results. We have started our journey with the goal of better livestock disease detection, a vital point of our research for mitigating the overall impact on the livestock industry. We have applied several machine-learning models for disease detection after utilizing a custom dataset with a train-test split of 70:30. Among the models used for disease detection, Neural Networks and Gradient Boosting Classifiers have shown the highest accuracy, around 90%. Besides, we have implemented different ensemble models, such as voting and stacking. This time stacking method provided an accuracy value of 92%. This achievement exceeds traditional models and underscores the effectiveness of detecting livestock diseases. Though the difference between ensemble and basic models is not so huge, we can say the data we used has fewer variations, and we haven't made huge improvements. As mentioned in the future goal, if we can collect more and more symptoms and diseases as an output class, we can have variations in our data. Then, o r models will perform better to detect the error. This creates an opportunity for stakeholders in the livestock industry to make informed decisions promptly. Integrating these machine learning models enriches computational efficiency and secures adaptability across various economic analysis sectors. In the future, we intend to expand our research exploration to find measures for better disease detection, early market analysis, and precise solutions for improvement in the livestock sector. Our research is crucial to livestock health and industry because it detects cattle diseases. It highlights various machine learning models that can favorably transform the livestock industry locally and globally more efficiently.

Bibliography

[1] M. Nasiru, U. Haruna, and A. Garba, "Economics of livestock marketing in gamawa local government area, bauchi state, nigeria," no. 304-2016-4818, p. 15, 2012. DOI: https://doi.org/10.22004/ag.econ.159412. [Online]. Available: http://ageconsearch.umn.edu/record/159412.

[2] B. Beyene, D. Hundie, and G. Gobena, "Assessment on dairy production system and its constraints in horoguduru wollega zone, western ethiopia," *Science, Technology and Arts Research Journal*, vol. 4, p. 215, Apr. 2016. DOI: 10.4314/star.v4i2.28.

[3] E. García-Gonzalo, Z. Fernández-Muñiz, P. J. Garcia Nieto, A. Sánchez, and M. Menéndez, "Hard-rock stability analysis for span design in entry-type excavations with learning classifiers," *Materials*, vol. 9, p. 531, Jun. 2016. DOI: 10.3390/ma9070531.

[4] C. Maltecca, D. Lu, C. Schillebeeckx, *et al.*, "Predicting growth and carcass traits in swine using microbiome data and machine learning algorithms," *Scientific Reports*, vol. 9, Apr. 2019. DOI: 10.1038/s41598-019-43031-x.

[5] A. Railey, F. Lankester, T. Lembo, R. Reeve, G. Shirima, and T. Marsh, "Enhancing livestock vaccination decision-making through rapid diagnostic testing," *World Development Perspectives*, vol. 16, p. 100 144, Nov. 2019. DOI: 10.1016/j.wdp.2019.100144.

[6] B. M. Tawheed, S. T. Masud, M. S. Islam, H. Arif, and S. Islam, "Application of machine learning techniques in the context of livestock," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 2029–2033. DOI: 10.1109/TENCON.2019.8929721.

[7] T. Bithari, S. Thapa, and H. K.C., "Predicting academic performance of engineering students using ensemble method," *Technical Journal*, vol. 2, pp. 89–98, Nov. 2020. DOI: 10.3126/tj.v2i1.32845.

[8] R. Guo, Z. Zhao, T. Wang, G. Liu, J. Zhao, and D. Gao, "Degradation state recognition of piston pump based on iceemdan and xgboost," *Applied Sciences*, vol. 10, p. 6593, Sep. 2020. DOI: 10.3390/app10186593.

[9] G. Limon, A. A. Gamawa, A. I. Ahmed, N. A. Lyons, and P. M. Beard, "Epidemiological characteristics and economic impact of lumpy skin disease, sheeppox and goatpox among subsistence farmers in northeast nigeria," *Frontiers in Veterinary Science*, vol. 7, Jan. 2020, ISSN: 2297-1769. DOI: 10.3389/fvets.2020.00008. [Online]. Available: http://dx.doi.org/10.3389/fvets.2020.00008.

[10] A. M. Musolf, E. R. Holzinger, J. D. Malley, and J. E. Bailey-Wilson, "What makes a good prediction? feature importance and beginning to open the black box of machine learning in genetics," *Human Genetics*, vol. 141, no. 9, pp. 1515–1528, Dec. 2021. DOI: 10.1007/s00439-021-02402-z.

[11] M. S. Rahman and G. Das, "Effect of covid-19 on the livestock sector in bangladesh and recommendations," *Journal of Agriculture and Food Research*, vol. 4, p. 100 128, Mar. 2021. DOI: 10.1016/j.jafr.2021.100128.

[12] R. F. Aunindita, M. Shiam Misbah, S. Bin Joy, M. A. Rahman, S. I. Mahabub, and J. Noor Mukta, "Use of machine learning and iot for monitoring and tracking of livestock," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, 2022, pp. 815–820. DOI: 10.1109/ICCIT57492.2022.10055766.

[13] G.-I. C, C. U, and A. O, "Rural livestock farmers perception of the role of information and communication technology tools in livestock production, management and improvement in imo state, nigeria," *South Asian Research Journal of Agriculture and Fisheries*, vol. 4, pp. 1–6, Feb. 2022. DOI: 10.36346/sarjaf.2022.v04i01.001.

[14] S. Fuentes, C. Gonzalez Viejo, E. Tongson, and F. Dunshea, "The livestock farming digital transformation: Implementation of new and emerging technologies using artificial intelligence," *Animal Health Research Reviews*, Jun. 2022. DOI: 10.1017/S1466252321000177.

[15] P. Geeks, *Gradient boosting algorithm in machine learning*, Oct. 2022. [Online]. Available: https://pythongeeks.org/gradient-boosting-algorithm-in-machine-learning/.

[16] J. Giordano, E. Sitko, C. Rial, M. Pérez, and G. Granados, "Symposium review: Use of multiple biological, management, and performance data for the design of targeted reproductive management strategies for dairy cows*," *Journal of Dairy Science*, vol. 105, no. 5, pp. 4669–4678, 2022.

[17] R. Roy, M. M. Baral, S. K. Pal, S. Kumar, S. Mukherjee, and B. Jana, "Discussing the present, past, and future of machine learning techniques in livestock farming: A systematic literature review," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1, 2022, pp. 179–183. DOI: 10.1109/COM-IT-CON54601.2022.9850749.

[18] J. You, D. Tulpan, M. C. Malpass, and J. L. Ellis, "Using machine learning regression models to predict the pellet quality of pelleted feeds," *Animal Feed Science and Technology*, vol. 293, p. 115 443, 2022.

[19] N. ABDALLAH, B. K. TEKELİOĞLU, K. KURŞUN, M. BAYLAN, and E. Ümit, "Vaccination and poultry (chicken) production," *Journal of Agriculture, Food, Environment and Animal Sciences*, vol. 4, no. 1, pp. 119–136, 2023.

[20] D. M. A. Bhamare, "Enhancing livestock management efficiency through operations research: Advancements in feed formulation, genetic optimization, and disease control," *International Journal of Engineering Applied Sciences and Technology*, vol. 8, no. 3, pp. 74–80, Jul. 2023.

[21] J. Gao, C. Bambrah, N. Parihar, *et al.*, *Building smart surveillance solutions for livestock farms*, Mar. 2023. DOI: 10.21203/rs.3.rs-2685214/v1.

[22] H. Hossain, M. Parvej, K. Brishty, *et al.*, "Epidemiological assessment of some infectious and non-infectious diseases and disorders of cattle and goat at certain milk-pocket area of sirajganj, bangladesh," *Veterinary Sciences: Research and Reviews*, vol. 9, Mar. 2023. DOI: 10.17582/journal.vsrr/2023/9.1.25.37.

[23] J. A. Hubbart, N. Blake, I. Holásková, D. Mata Padrino, M. Walker, and M. Wilson, "Challenges in sustainable beef cattle production: A subset of needed advancements," *Challenges*, vol. 14, no. 1, p. 14, 2023.

[24] l. I.D, O. Daodu, A. J.O, *et al.*, "Review of the major microbial diseases associated with high mortality in ruminants in nigeria," *Nigerian Journal of Pure and Applied Sciences*, pp. 4592–4606, Dec. 2023. DOI: 10.48198/NJPAS/23.A22.

[25] M. Islam, S. Zabeen, F. Rahman, M. Islam, and F. Kibria, "Analysis of uncertainty in different neural network structures using monte carlo dropout," Ph.D. dissertation, Jan. 2023.

[26] M. Jamil, S. Sidra1, A. Hussain, M. Imran, and A. A. Sheikh, "Assessment of air quality of livestock farms in district kasur and lahore," pp. 1–8, Jun. 2023. DOI: https://dx.doi.org/10.17582/journal.pjz/20230224040218.

[27] J. Marwah, K. Kanwar, Sonia, P. Vaidya, and G. Gupta, "An analysis on animal welfare, production and cost through statistical and machine learning techniques," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2023, pp. 1853–1857. DOI: 10.1109/ICACITE57410.2023.10183050.

[28] D. Mwiti, *A comprehensive guide to ensemble learning: What exactly do you need to know*, Aug. 2023. [Online]. Available: https://neptune.ai/blog/ensemble-learning-guide%7D.

[29] Y. Pan, Y. Zhang, X. Wang, X. X. Gao, and Z. Hou, "Low-cost livestock sorting information management system based on deep learning," *Artificial Intelligence in Agriculture*, 2023.

[30] K. Patel, "Economic losses due to important diseases of dairy bovines in sabarkantha district of gujarat," Sep. 2023.

[31] I. Peak, *Livestock symptoms and diseases*, May 2023. [Online]. Available: https://www.kaggle.com/datasets/researcher1548/livestock-symptoms-and-diseases.

[32] A. Shirzadifar, G. Plastow, J. Basarab, *et al.*, "397. a machine learning approach for predicting the most and the least feed-efficient groups in beef cattle," Feb. 2023, pp. 1656–1659. DOI: 10.3920/978-90-8686-940-4_397.

[33] M. E. Suaza-Medina, F. J. Zarazaga-Soria, J. Pinilla-Lopez, F. J. Lopez-Pellicer, and J. Lacasta, "Effects of data time lag in a decision-making system using machine learning for pork price prediction," *Neural Computing and Applications*, 2023.

[34] D. R. Yehoshua, *Random forests*, Jul. 2023. [Online]. Available: https://medium.com/@roiyeho/random-forests-98892261dc49.

[35] [Online]. Available: https://www.knime.com/blog/a-friendly-introduction-to-deep-neural-networks.

[36] Ashtired11, *Cattle diseases*, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/ashtired11/cattle-diseases.

[37] J. Campbell, *Bacterial pneumonia in cattle with bovine respiratory disease complex - respiratory system*. [Online]. Available: https://www.msdvetmanual.com / respiratory - system / bovine - respiratory - disease - complex / bacterial - pneumonia-in-cattle-with-bovine-respiratory-disease-complex.