Identification and Modeling of Significant External Factors Causing Bovine Mastitis in Cows using ML Methods

By

MALIHA IQBAL

Master of Science in Bioinformatics (Fall 2020)

2020-MS BI-5-00000326862

Supervisor

Dr.Zamir Hussain

School of Interdisciplinary Engineering and Sciences (SINES)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

October, 2022

Declaration

I, Maliha Iqbal declare that this thesis titled "Identification and Modeling of Significant External Factors Causing Bovine Mastitis in Cows using ML Methods" for the degree of Master of Science in Bioinformatics has been produced entirely by me, during the scheduled period of study. The experimental work is exclusively mine, with clear indications of and acknowledgements for any joint contributions. My research has not been submitted for any previous degree.

Maliha Iqbal

2020-MS BI-5-00000326862

School of Interdisciplinary Engineering and Sciences (SINES)

National University of Sciences and Technology (NUST)

Copyright Notice

Dedication

To my family, teachers and friends

Acknowledgments

In the name of Allah, the Most Gracious and the Most Merciful, all praise to Him for His strength and blessings that helped me to complete my research. I would first like to express my sincere gratitude to my supervisor, Dr. Zamir Hussain, who has supported, encouraged, and assisted me greatly during this research project. His tremendous assistance with helpful criticism and recommendations throughout the research and thesis preparation have helped this study succeed. I would like to thank NUST and I value the research environment that SINES has provided. I would like to extend my thanks to Dr. Muhammad Tariq Saeed and Dr. Zartasha Mustansar, members of my Guidance and Examination Committee (GEC), for always helping me while I was conducting my research. I want to thank the Principal, and the Head of Department of SINES. I acknowledge all my teachers, who guided me at every step. Special thanks to my beloved parents for their unconditional love and prayers. Lastly, I would like to thank all of my friends who always encouraged and helped me.

List of Abbreviations

SFMT .... Surf Field Mastitis Test

SPSS .... Statistical Package for Social Sciences

AI .... Artificial Intellegance

SVM .... Support Vector Machines

ANN .... Artifical Neural Networks

LDA .... Linear Discriminant Analysis

RF .... Random Forest

ML .... Machine Learning

KNN .... K Nearest Neighbors

NB .... Naïve Bayes

Contents

List of Figures

List of Tables

Abstract

Bovine mastitis is a condition in which mammary gland has inflammation and the major causative agent is bacterial infection. Microbes usually cause mastitis when they enter teat through teat canal. Streptococci, staphylococci, and gram-negative rods cause most infections. It is an important health issue in dairy farms. Cow welfare and longevity are affected, that can lead to economic losses due to reduced production of milk, poor quality of milk and cost of treatment. The enhancement of herd hygiene conditions is one of several options available today to attain and maintain a good udder health status for dairy cows. To increase an animal's resistance to mastitis, enhanced mastitis detection methods and genetic selection of animals are used. Mastitis is one of the most frequently occurring and costly disease of dairy cows worldwide. US dairy industry alone has estimated annual cost of US 2 billion dollars. Various measures should be implemented to ensure the health and well-being of the animals that are the foundation of dairy industry. Mastitis is classified as clinical and subclinical type depending on the manifestation of the disease. Clinical mastitis occurs as an inflammatory response to infection and there is visibly abnormal milk with chemical and physical changes. The mammary gland also may exhibit change in its morphology. The subclinical form occurs when both milk and udder appear normal without noticeable manifestations of inflammation. Subclinical mastitis is more prevalent than clinical mastitis and causes the greatest overall losses in most dairy herds worldwide. In general, the economic loss is estimated to be approximately 100 Euros per cow.

It is thus imperative to understand the risk factors that have high association with mastitis. This information can then be used to target those high-value factors to deliver the best impact for prevention strategies. In order to achieve the aforementioned targets a study has been conducted in which 432 cows are randomly selected from 40 farms in the Rawalpindi and Islamabad. Risk factors that could have a significant impact on mastitis development are selected from the literature for a total of 28 categorical factors. Of these, 18 factors are herd-level and 10 are animal-level. The Surf Field Mastitis (SFMT) test is used in order to classify the diseased vs. normal cases.

The purpose of this study is identification of significant external factors causing bovine mastitis in cows considering local data set of 432 instances keeping in view the issue of multi-collinearity. Our major target is the identification of factors that could differentiate between diseased and normal cow with efficient and acceptable accuracy. Secondly, development of predictive models using different machine learning

methods like K nearest neighbor (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) considering a binary dependent variable (state of disease either yes or no) and 28 external factors. After short-listing factors on basis of chi-square analysis, the factors based on assessment of clinical professionals are selected.

The combinations include all set of factors, and then reduced subsets of 10 factors based on chi-square analysis. 14 factors based on assessment of clinical professionals and 7 based on union of statistically selected factors and subject knowledge. SVM performed best based on 28 factors with 70 percent sensitivity while on the basis of 10, 14 and 7 factors ANN performed best as compared to the other models with 65 percent sensitivity with respect to 10 factors and 70 percent sensitivity with respect to 7 and 14 factors. For model validation, 10-fold cross validation scheme is used. Based on the provided details, the study recommends the use of ANN with 7 factors named factors named feed sharing, washing of udder, condition of udder, dipping (pre post teat dipping), last (milking the mastitis cow last), use of hormones and udder hygiene score to predict category of the target class.

Chapter 1

Introduction

The gut microbiome, bone mass, and cardiovascular health have all shown to take benefit from dairy product consumption. Milk is a nutrient-dense food especially crucial for young toddlers and newborns who require nutrition and foods with lots of energy for development of the mind and body. According to research, drinking milk helps prevent stunting [1]. Small farmers own the majority of the livestock, and large production is discouraged due to high transportation costs, limited infrastructure, and other factors [2]. Farmers struggle to maintain hygiene due to a lack of knowledge and resources, which causes infectious diseases in animals including tuberculosis, Brucellosis, mastitis, etc. Mastitis is the second most commonly occurring disease among dairy cows, making it a critical issue for the dairy industry [3].

1.1 Mastitis

Bovine mastitis, is the "infection of the mammary gland", have an infectious or noninfectious etiology. Organisms such as bacteria, mycoplasma, yeasts and algae are identified as causes of the disease. 137 distinct organisms are known to be the cause of mastitis. Majority of mastitis in the United Kingdom has bacterial origin and simply 5 species of bacteria (Escherichia coli, Streptococcus uberis, Staphylococcus aureus, Streptococcus dysgalactiae and Streptococcus agalactiae) results 80 percent of the disease. Mastitis is classified as clinical and subclinical type depending on the manifestation of the disease. Clinical mastitis have visible symptoms whereas subclinical mastitis does not show visible symptoms the milk appears normal but usually has an elevated somatic cell count. Cow welfare and longevity are affected, that can lead to economic losses due to reduced production of milk, poor quality of milk and cost of treatment. Early detection of mastitis is imperative [4].

1.2 Classification Of Mastitis

Based on the source of infections, mastitis can be split into two groups:

1- Contagious

2- Environmental

1.2.1 Contagious Mastitis

Infections, which are spread from cow to cow during the milking process. The most contagious pathogens causing intramammary inflammation are S. aureus, Str. agalactiae, and Str. Uberis [5].

1.2.2 Environmental infections

Infections, which are spread by bacteria in the cow's environment. The environmental pathogens causing mastitis in cows are E. coli, Klebsiella (K.) pneumoniae, Enterobacter aerogenes, and Str. Uberis [5].

1.3 Types of Mastitis

The two main types of mastitis are

1- Clinical mastitis

2- Subclinical mastitis

1.3.1 Clinical Mastitis

It is associated with inflammatory response to infection that causes abnormal milk and when the inflammation increases, changes within the udder might be apparent. Clinical cases with less apparent signs are termed as moderate and mild. The inflammatory response which, includes general involvement of (fever, anorexia, shock), then this case is termed severe. If the onset is extremely fast, it usually happens with severe clinical cases [6]. The major pathogens that are involved in clinical mastitis are Escherichia Coli, Staphylococcus Aureus and Streptococcus Uberis [7].

Risk factors associated with clinical mastitis

Retained placenta, female internal reproductive organ infections, pyometra, dystocia, also as twin births were among the diseases studied. Throughout the prepubertal phase, such abnormalities were found to be linked to clinical mastitis [8].

1.3.2 Sub-Clinical Mastitis

Infection exists without apparent signs of local inflammation or systemic involvement. Though abnormal milk may appear, it is usually asymptomatic. Chronic infection is when the infection remains for two months. Once developed, most of these infections remains for entire lactations or the whole life, but this varies with the causative pathogen [6]. The Sub-clinical mastitis is caused by minor pathogens such as Staphlococcus Aureus and Corynebacterium Bovis [7].

Risk factors associated with Sub-clinical mastitis

1- Age

2- Pendulous Udders

3- Dirty Udders

4- Teat End Lesions

5- Milking diseased cows before milking healthy cows.

6- Moving heifers to restricted housing on the day of calving [8].

1.4 Prevalence Of Mastitis

The average annual incidence of clinical mastitis was discovered to be between 25 and 30 occurrences per 100 cows. The average economic death toll from mastitis is thought to be roughly 150 Euros per cow per year. Mastitis is 16.72 percent prevalent in Pakistan. In Punjab, the annual total of privations caused by clinical mastitis was roughly 240 million two decades ago. [9].

1.4.1 Prevalence of mastitis at International level

Internationally this disease caused 35 billion US Dollor damage annually. In the United States, mastitis is thought to result in annual economic losses of 2 billion US Dollor [9]. In India, clinical and subclinical mastitis reduced milk output by 50 percent and 17.5 percent, respectively, according to a survey conducted about 20 years ago [10].

1.4.2 Prevalence of mastitis at National level

Mastitis has a much higher prevalence rate all over the world, including Pakistan, with a prevalence of 16.72 percent. Although it was predicted two decades ago that overall losses caused by clinical mastitis in Punjab province are estimated to be Rs. 240 million per year, information on current losses due to this illness are not accessible in Pakistan.[10]. Previous research in other parts of Pakistan shows that the prevalence of clinical mastitis in buffaloes with cattle was 21.08 and 16.72 percent, correspondingly [10].

1.5 Risk Factors

Risk Factors for Bovine mastitis can be external as well as internal.

1.5.1 External Factors

Some of the external factors are:

1- Floor Type

2- Milking Hygiene

3- Udder injury

4- Tick infestation of the udder [10]

1.5.2 Internal Factors

Some of the internal risk factors are:

1- Parity

2- Stage of Lactatio

3- History of Mastitis

4- Breed [10]

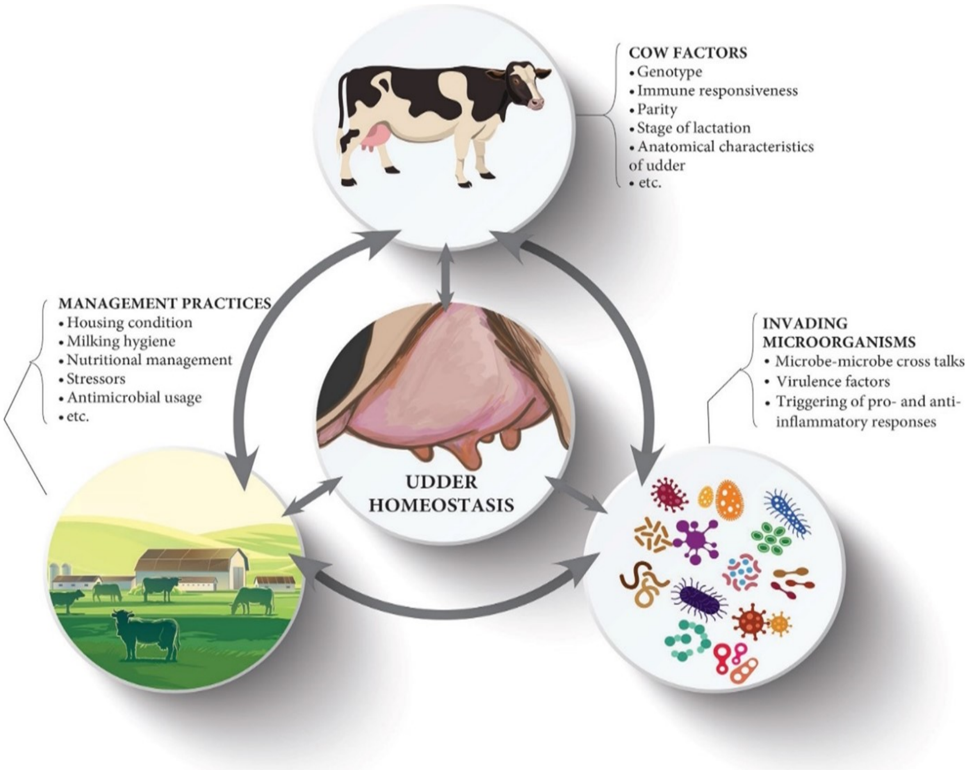

Figure 1.1: Udder Homeostasis

1.6 Symptoms

The most obvious symptoms of clinical mastitis are:

1. The udder will swell and appear red. It will also feel extremely warm to the touch.

2. Touching the cow's udder will cause discomfort to the cow.

3- In extreme situations, the cow's body temperature will rise, her milk may seem watery and contain flakes, clots, pus, or blood, and her body temperature will rise [11].

1.7 Diagnosis

For diagnostics some of the tests that can be used are:

1. Surf Field Mastitis Test

2. California Mastitis Test

3. Electrical Conductivity Tests

4. Somatic Cell Count Measurement

The diagnostic test used in this study is:

1.7.1 Surf Field Mastitis Test

Make a 3 percent Surf solution. Add 6 teaspoons of Surf to 0.5 liters of water, mix, filter and heat the solution. Take the milk, add an equal volume of 3 percent solution, stir this mixture for 30 minutes and then check for precipitate or gel formation (for mastitis) [12].

1.8 Prevention and Treatment

We can prevent new infections by focusing on management efforts by reduction of the presence of pathogens on teat ends. Clean bedding, hygienic udders at the time of milking, avoiding the use of water during the milking process (except for maintaining hygienic milking units), and maintaining teat-end health all have a positive impact on controlling the disease. Contagious pathogens, can be dealt with an effective germicide as a post milking teat dip for preventing the transmission of disease. Other things that can help in prevention of the disease are use of clean and dry towels for teats, milkers should wear gloves, germicide should be used before milking, and cow should not be under any stress of over milking, the milking equipments should be cleaned if they are exposed to an infected cow. If it is possible, the infected cow should be segregated from the healthy ones. Infections caused by environmental pathogens can be reduced by good milking routine.

Things that have a positive effect on environmental mastitis control include:

1- Clean Bedding

2- Reduction of Heat Stress

3- Udder hair removal

4- Reducing udder swelling by nutritional management of potassium and sodium intake

5- Avoiding water surrounded areas

6- Stalls maintained properly

7- Prevention of frostbite and exposure to fly [13]

For treatment the options are, intramammary antibiotics, classic mastitis tubes, and systemic antibiotics administered intramuscularly or subcutaneously. Intramammary antibiotics should be the first-line treatment for cows with mild, uncomplicated mastitis in the first quarter [13].

1.9 Problem Statement

Keeping in view the details provided in the literature review section, following are few major issues with respect to the identification and modeling of significant external factors causing bovine mastitis.

1- Disparity exists in the reported literature with respect to factors causing the disease

2- Presence of multi-collinearity between different independent factors hence limiting the usefulness of estimates provided through modeling

1.10 Proposed Solution

This study would provide two steps, for the standardization and dealing with multicollinearity between factors. The first step would help in identification of significant external factors causing the disease. In the second step, modeling of shortlisted factors is done using machine learning methods like K nearest neighbor (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) keeping in view the issue of multi-collinearity.

1.11 Advantages

This study has a great deal of potential to help non-specialist veterinary clinicians quickly identify the disease at the herd level and swiftly implement effective control measures for a disease that is extremely detrimental to animal health, productivity, welfare, and the use of antibiotics. Machine learning analytics can enhance monitoring techniques and assist dairy farmers and staff in proactively identifying cows that might be diseased.

1.12 Objectives

Following are the main objectives of this study:

1. Identification of significant external factors causing bovine mastitis in cows considering local data set of 432 instances keeping in view the issue of multi-collinearity. Our major target would be that the identified factors could differentiate between diseased and normal cow with efficient and acceptable accuracy.

2. Development of predictive models using different machine learning methods like KNN, SVM and ANN considering a binary dependent variable (state of disease either yes or no) and 28 external factors.

Chapter 2

Literature Review

This chapter describes various studies conducted at the international and national level on the important internal and external factors associated with the prevalence of mastitis disease.

2.1 International Studies

Tania Bobbo et al. examined eight alternative machine-learning models, including LDA, NN, GLM, SVM, RF, CART, KNN, and NB, for the prediction of udder health status based on somatic cell count. All of the models displayed prediction accuracy of at least 75 percent. The most effective approaches were NN, RF, and linear. The cows were from commercial herds and had no invasive treatments. Data was gathered between January 2018 and January 2020, and milk samples were taken as part of standard milk collecting procedures. The dataset contains information about the herd, the cow (ID, breed, lactation stage, and number of eggs), the sample date, the daily milk yield (kg/day), the milk composition (fat, protein, casein, and lactose levels, pH, and urea (mg/100 ml)), and the daily milk yield (kg/day). In this investigation, a dataset with 18,442 records was employed. They utilised 15 characteristics to train 8 ML models. The top five features out of these 15 traits include casein, protein, differential somatic cell count, protein and lactation stage in cows from the same herd and day. Several metrics, including true and false positive rates, precision, F-score, and accuracy, were used to assess how well the ML approach performed on the test and validation sets. One of the most frequently used metrics is accuracy, which is the proportion of true forecasts to all other predictions. However, accuracy by itself is insufficient to assess the effectiveness of a model. As a result, other metrics, such as recall or sensitivity and precision, were also used in the evaluation of the model. In order to find the model that best balances both false positive. This was most likely brought on by sensitivity to noise. In particular, the linear technique was among the lowest false-positive errors, NN and RF were among the lowest false-negative errors, and the best four methods had the lowest overall prediction error rate on the validation set. The total error rate for the CART approach was good, comparable to that of NN, RF, LDA, and GLM. The lowest false negatives were found with NB, but the largest false positive errors. The most accurate techniques for predicting udder health seemed to be NN, RF, and linear approaches. Even while some strategies outperformed others, the overall forecast accuracy was still higher than 75 percent. Based on the findings, machine-learning algorithms appear to be useful tools for enhancing farmers' decision-making. The day after testing, machine learning analytics can enhance monitoring techniques and assist farmers in proactively identifying cows that might have high somatic cell counts. [14].

Robert M. Hyde et.al used RF to train 20 features for herd level diagnoses of mastitis in dairy herds. It is important to understand the infections' mechanism of transmission because mastitis can be both infectious and environmental. Based on the transfer of infections, environmental mastitis can be further classified into environmental lactation period (EL) and environmental dry period (EDP). Between 2009 and 2014, data from 1000 farms was gathered. After filtering and preprocessing, a total of 278 farms and 290 farms were available for cross-validation (CV) and external validation (EV) sets, respectively (CONT vs ENV). For the diagnosis of CONT vs. ENV, Random Forest demonstrated accuracy of 98 percent, and for the diagnosis of EDP vs. EL, accuracy of 78 percent. Accuracy, positive predictive value (PPV), and negative predictive value (NPV) were assessed as model performance metrics after each stage of model tuning and feature engineering.[15].

To determine the frequency and risk factors for subclinical mastitis in dairy cows kept by smallholder farmers in Tanzania's Iringa and Tanga regions, E. D. Karimuribo et al. conducted a study on 200 randomly chosen farms on each of those two regions' respective islands. The California Mastitis Test (CMT) was used in this investigation to identify subclinical mastitis, and bacterial cultures were obtained from 1500 milk samples from 434 clinically healthy cows. Having CMT results, there were 75.9 percent (46.2 percent) of cows (and quarters) with asymptomatic mastitis. However, culture findings showed that there were 43.8 percent (24.3 percent) of cows (and quarters) with subclinical mastitis. Age, peak milk yield, imported cattle (as opposed to homebred), and the boreane breed were all strongly linked to an increased chance of a CMT-positive quarter. A considerably decreased incidence of CMT-positive quarters was linked to manual milking using a stripping method. Culture-positive cows were more likely to be found in CMT-positive cows, as were older and store-bought cows. [16].

Liliana Fadul-Pacheco et.al used various machine learning classification methods (Naive Bayes, Random Forest, Extreme Gradient Boosting) to identify clinical mastitis (CM) cows during their first lactation (1st lactation) The development of CM was (continuously) recognised and predicted daily. Data from other software programmes was merged and supplied into the algorithm. Using the random forest algorithm, the best predictions were made in both situations. With the original lactation model and the continuation model, the system accurately categorised 71 percent and 85 percent of CM cows, respectively. The precision for both studies was 72 percent. The findings demonstrate that different data streams can be combined to create decision support systems that are both predictive and prescriptive. Short-, medium-, and long-term decision-making may be aided by having two distinct algorithms forecast direct risk and overall risk throughout the first lactation while operating simultaneously. [17].

2.2 National Studies

Asghar Khan Et.al conducted this study to estimate the prevalence and assess potential risk factors for asymptomatic mastitis in lactating buffalo in the Pothohar region of Pakistan. It is believed that a significant barrier to the growth of Pakistan's dairy industry is asymptomatic mastitis. One of the top nations with the highest prevalence of subclinical mastitis is Pakistan. In the Pothohar region of the Rawalpindi district in Punjab, 30 commercial and subsistence farms were therefore the subject of a cross-sectional questionnaire-based survey. Random samples were taken from 196 dairy cows. Face-to-face interviews with farmers, managers, and owners allowed for the collection of information on a number of health, management, and biosecurity variables. California mastitis testing was performed on milk samples that had been collected. Epi-Data handled the data entry and validation. SPSS was used for data analysis. Regression analysis and chi-square were used. Overall frequency was found to be 67.3 percent. [18].

Riaz Hussain et.al conducted study on 453 lactating cows of different breeds on 21 farms. Milk samples were taken from these animals and mastitis was diagnosed using the California Mastitis Test (CMT). Epidemiological data on animals and management were gathered, examined, and insightful conclusions were made. The findings of the t-test revealed a significant relationship between the target variable and body weight, udder depth, teat length, teat diameter, and distance between teat base and floor (P 0.01). Lactation (P 0.0001), teat-to-floor distance, number of eggs laid, udder shape, teat shape, live weight, use of oxytocin, lactation system, and milk leakage were all found to be significantly associated with mastitis in bivariate frequency analysis.. Logistic regression research revealed a substantial positive correlation between mastitis-affected teat, teat diameters (apex, central, base), milk yield, and udder shape, but a significant negative correlation between teat length, frequency of culling, and number of attendees. Milk yield, teat and/or udder pathology, live weight, suspended udder and feeding system, udder depth, and teat shape. Our findings suggest that a number of risk factors are closely linked to cattle mastitis. [19]. This study was done by M.Q. Bilal et al. to determine whether factors, including peri-urban and urban locations, affect the occurrence of mastitis in buffaloes in the Faislabad region. Based on a questionnaire survey, the data. Numerous criteria were covered in the questions, such as the state of the milk from the affected teats, the number of animals with swollen/reddened teat quarters, the soil conditions, the manner of milking, etc. According to a study, rural areas had a higher prevalence of mastitis than urban areas, with 25.12 percent and 19.74 percent, respectively. Between 4 and 6 months following calving, the highest prevalence was noted in both rural (45.08 percent) and suburban (45.76 percent) locations. Mastitis can benefit from cement floors.[9].

Considering the aforementioned details of the published literature, major gaps include, disparity exists in the reported literature with respect to factors causing the disease and presence of multi-collinearity between different independent factors hence limiting the usefulness of estimates provided through modeling. This study would provide two steps, for the standardization and dealing with multicollinearity between factors. The first step would help in identification of significant external factors causing the disease. In the second step, modeling of shortlisted factors is done using machine learning methods like K nearest neighbor (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) keeping in view the issue of multi-collinearity.

Chapter 3

Methodology

The aim of this study is to identify important external characteristics of mastitis for early screening of cattle for mastitis disease by various statistical and machine learning approaches. In this study, we performed the following steps for the development of the predictive models.

1. Dataset
2. Data Preprocessing
3. Selection of Attributes
4. Model Development
5. Assessment Analysis

Details of these steps are discussed below. Various software such as SPSS, Excel, and Anaconda are used to analyze the data and develop the models described in this chapter. The data is in the form of classes and labels. Categorical data is further divided in the context of attributes into nominal and ordinal data. Some attributes are nominal in nature, while others are ordinal in nature. In our study, age is ordinal in nature, whereas all other 27 factors listed in table 3.1 are nominal in nature.

A complete workflow of the proposed methodology is presented in figure 3.1.

3.1 Dataset

A secondary dataset having dimensions of 432(rows) x 29(columns) is used in this research. Here rows represent instances, i.e. cows and columns represents factors (29 in total) observed on each instance treated as 28 independent variables and one dependent variable while modeling. Data is provided by

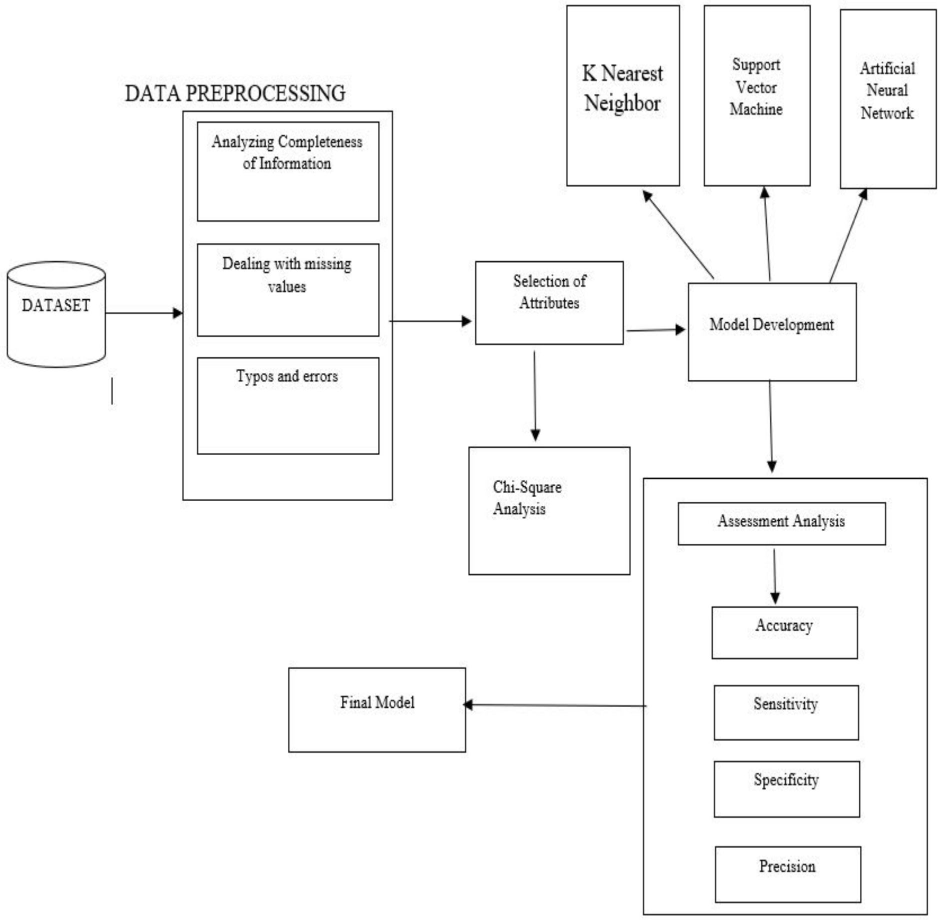

Figure 3.1: Methodology Flowchart

the Anti Bacter Research Group of ASAB (Atta-Ur-Rahman School of Applied Biosciences, National University of Sciences and Technology (NUST), Pakistan) (http://asab.nust.edu.pk/research/antibacter-research-group/). The respective group collected the data via clinical examination of the udder and teats of each cow by visiting 40 dairy farms of Rawalpindi and Islamabad. A structured questionnaire consisting of 28 statements covering qualitative information of risk factors is used to collect information from the farm owners and/or associated staff. These risk factors are mainly driven through literature. For the collection of data, the complete procedure is carried out in accordance with relevant guidelines and regulations. An institutional review board in ASAB was formed that approved the experimental protocols being followed during the collection of data and ensuring no privacy or ethical violations of collected information. The composed data was then screened through Surf Field Mastitis Test (SFMT) to check clinical or subclinical levels. Description of 28 external factors is listed in Table 3.1. Notable points are

a. Factors treated as independent variables are categorical in nature with majority of them having binary category.

b. We have a binary dependent variable with 0 being normal and 1 being disease (mastitis)

3.2 Data Preprocessing

The process of converting raw data into a format that computers and machine learning can comprehend and evaluate is known as data preprocessing. Outlier identification, estimation of missing values, feature selection, etc. are all included. Real raw data, such as text, photographs, videos, etc is disorganized. They are frequently incomplete, lack a regular and consistent design, and not only have the potential to contain faults and contradictions. Machines prefer to deal with accurate and well-organized data, reading information as 1s and 0s. Calculating structured data, such as integers and percentages, is made simple by this. But prior to analysis, unstructured material in the form of text and images must be prepped and formatted.

In this study, data preprocessing entails assessing and verifying that each factor's information is full, for as by addressing typographical errors and missing values.

Table 3.1: Description of 28 External Factors

Sr.No	**Factors**	**Categories**	**Description of categories**
1.	Management System	Two	Intensive \ Semi intensive [1]
2.	Bedding Material	Two	Yes \ No
3.	Housing	Two	Group \ Stall [2]
4.	Floor Type	Three	Muddy \ Concrete \ Mixed
5.	Milking Method	Two	Manual \ Machine
6.	Milking Routine	Two	Two times \ Three times
7.	Washing of Udder	Two	Whole udder \ Only teats[3]
8.	Drying of Udder	Two	Yes \ No
9.	Position of Udder	Two	Normal \ Pendulous[4]
10.	Condition of Udder	Three	Atrophy[5] \ Normal \ Swelling
11.	Presence of Ticks	Two	Yes \ No
12.	Lesions (teat end lesions)	Two	Yes \ No
13.	Dipping (pre \ post teat dipping)	Two	Yes \ No
14.	Standing Position after Milking	Two	Yes \ No
15.	Last (milking the mastitis cow last)	Two	Yes \ No
16.	Use of Hormones	Two	Yes \ No
17.	Use of Towel	Two	Yes \ No
18.	History of Mastitis	Two	Yes \ No
19.	Udder Hygiene Score	Three	Moderately dirty \ Slightly dirty \ Very dirty
20.	Feed Sharing	Two	Yes \ No Urban
21.	Manure Removal	Two	Daily \ Once a week
22.	Number of Attendees	Two	Only one \ Two people
23.	Location of Farm	Two	Rural \ Urban
24.	Size of Herd	Two	10 no. of cows \ 10 no. of cows
25.	Type of Herd	Two	Mixed \ Single type
26.	Age	Two	5 years \ 5 year
27.	Breed	Two	Local \ Cross[6]
28.	Stage of Lactation	Three	Early \ Mid \ Late[7]

Note: 1: Intensive: Cows are not allowed to go out, Extensive: Cows allowed to go out 2:Stall: For rest, feeding, milking, and watering, each cow is tied up separately in a stall. 3:Udder: A sizable organ with a bag-like form that houses two or more milk glands, each of which drains into a distinct teat on the lower surface 4:Pendulous: Pendulous udders are a common problem among high-producing dairy cows. Udder is tilted to one side 5:Atrophy: The size of an organ decreases as a result of disease. 6:Cross: Combination of these different breeds 7:Early: 1-2 months after milking mastitis occurs ,Mid:3 months after lactation mastitis occurs, Late :lactation ends then mastitis occurs.

3.3 Selection of Attributes

Choosing the most crucial qualities, features, categories, etc. for the analysis is known as the selection of attributes. The ML models are trained using these factors. It is vital to keep in mind that the training process will take longer and potentially produce less accurate results if there will be more factors. This is due to the possibility that some factors in the data have duplicate features or no longer exist. In this study, we will use two-step approach. In the first step, association between independent factors will be determined using Chi square analysis. In the second step, the association of shortlisted external factors with prevalence of mastitis will be determined. The confidence interval taken in order to select the factors is 95 percent. And as a result of this analysis factors are selected on the basis of P-value.

3.3.1 Chi-Square Analysis

Chi-square test is also known as Pearson's chi-square. The standard Chi-square statistics is defined as

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

It helps us to determine that whether the difference between observed and expected values is due to chance or due to existing relationship between factors being studied. Hence, a chi-square test is one of the best choice to help us understand and interpret the relationship that exists between two categorical variables. In the output table of chi-square analysis, the asymptotic significance, or p-value, is the value that determines the statistical significance of the relationship. In all tests of significance, if p ¡ 0.05, we can say that there is a statistically significant relationship between the two factors. And if pearson chi-square value is higher between two factors this shows greater association between the two factors.

3.4 Model Development

Predictive models like KNN, SVM and ANN are developed in order to determine the strongest association of independent factors (external factors) with dependent variable (disease).

3.4.1 KNN

The KNN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories. The kNN algorithm determines the separation in the training sample and test sample in data, then returns the k examples that are the closest together. It is guaranteed to locate the precise k nearest neighbours and has a linear time complexity [20]. A new data point is classified using the KNN algorithm based on similarity after all the existing data has been stored. This means that utilizing the K-NN method, fresh data can be quickly and accurately sorted into a suitable category. Although the K-NN approach is most frequently employed for classification problems, it can also be utilised for regression [21]. Since KNN is a non-parametric technique, it makes no assumptions about the underlying data. It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it uses the dataset to perform an action when classifying data. The KNN method simply saves the information during the training phase, and when it receives new data, it categorizes it into a category that is quite similar to the new data [22].

3.4.2 SVM

A supervised machine learning model called a support vector machine (SVM) employs classification techniques to solve two-group classification problems. An SVM model can classify new text after being given sets of labelled training data for each category [23]. They offer two key advantages over more recent algorithms like neural networks: greater speed and improved performance with fewer samples (in the thousands). As a result, the approach is excellent for text classification issues, where it has only have access to a dataset with a few thousand tags on each sample. The SVM modelling technique creates a separating hyperplane with the maximum margin in order to operate geometrically. On mildly skewed data, SVM is more accurate when compared to other conventional classifiers. Because only SVs are utilized for classification, a large number of majority samples that are outside of the decision boundary can be eliminated without having any impact on the classification. The classification performance on the positive class will suffer if an SVM classifier is sensitive to a high class imbalance. It has a high propensity to produce classifiers with severe estimation biases toward the majority class, leading to a high proportion of false negatives [24].

3.4.3 ANN

An ANN is a collection of connected input and output networks where each connection has a weight, few layers, one input layer, and one layer of output. In order to execute neural network learning, the weight of connection is adjusted. The performance of the network is enhanced by adjusting the weight iteratively. Depending on the connectivity, ANN can be divided into two groups: feed-forward recurrent network [25]. ANN compares the processed output of the network—often a prediction—against the desired output. The error is in this discrepancy. The network then modifies its weighted associations using this error value and a learning strategy. The neural network will produce output that is increasingly comparable to the goal output as modifications are made over time [26].

3.4.4 Train-Test-Split

The train test split method in Sklearn model selection divides data arrays into training data and testing data subsets. The model is trained using a training set, evaluated using metrics from the test set, and the procedure is then carried out k times. The process is then carried out once more until learning stabilizes and stops progressing [27]. By default, Sklearn train test split divides the two subsets into random groups [28]. In our study, the dataset was divided into 80 percent training and 20 percent test set. Flowchart of train-test split is shown in figure 3.2

3.5 Assessment Analysis

Assessment analysis of the developed models is determined by using accuracy, sensitivity, specificity and precision.

3.5.1 Accuracy

The percentage of classifications a model successfully predicts divided by the total number of predictions is known as model accuracy.

3.5.2 Precision

When true positives are divided by total number of positive predictions this is known as precision.

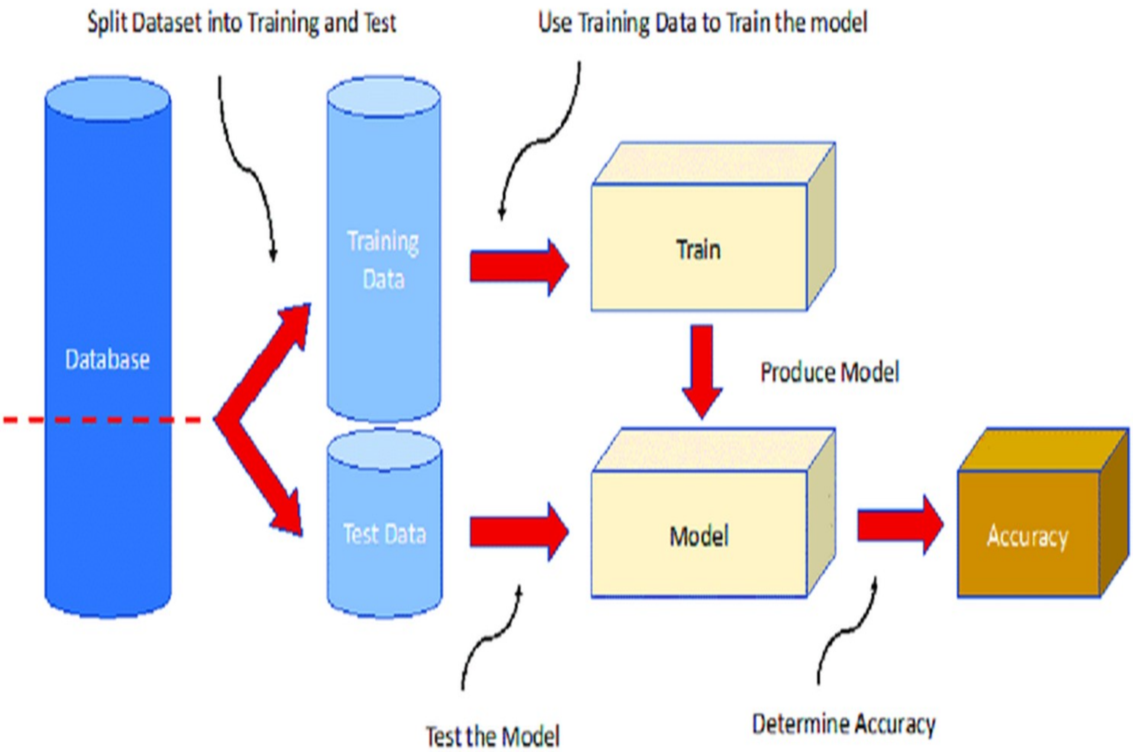

Figure 3.2: Flowchart of tain-test split

3.5.3 Sensitivity

How accurately machine-learning model can identify positive instances is measured by its sensitivity. The true positive rate (TPR) or recall are other names for it.

3.5.4 Specificity

The percentage of true negatives that the model successfully predicts is known as specificity.

3.5.5 10-fold Cross-Validation

A statistical technique known as cross-validation compares and evaluates learning algorithms by splitting data into two parts: one for learning or training a model and the other for model validation. The training and validation sets must overlap in subsequent rounds during a typical cross-validation so that every data point gets a chance to be validated against. K-fold cross-validation is the most fundamental type of cross-validation. Other cross-validation techniques are variations on k-fold cross-validation or use many iterations of k-fold cross-validation [29].

We divide the dataset into 10 parts using 10-fold cross validation; 9 of the 10 parts train the classifier, and the data from the training phase is used to validate (or test) the 10th part. This process is repeated 10 times, so that at the conclusion of the training and testing phases, each part has been used as both training and testing data [30].

Chapter 4

Results

4.1 Data Preprocessing

Data preprocessing is a process of converting raw data into machine-readable form. In our primary dataset, we have three such kind of factors, which contains the typos errors. The name of these factors are size of herd, age and number of attendees. (Table 4.1) According to the information provided in the table, 4.1 there are 17 percent of the values in the factor named herd size contains the typos errors while 48 percent and 69 percent of the values in the factor named age and no of attendees contain typos errors respectively.

4.2 Selection of Attributes

4.2.1 Association of factors with target variable

We checked the association of independent factors with target variable. For this propose we used chi-square analysis. Chi-square analysis is also used to check the association between categorical factors since our primary dataset has categorical independent factors and categorical dependent variable therefore, we

Table 4.1: Data Preprocessing

Sr.No.	**Factors**	**No of Typos Errors**	**Percentage**
1	Size of Herd	73	17
2	Age	209	48
3	Number of Attendees	297	69

Table 4.2: Association of Breed with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Management System	1.724	0.189
2	Bedding Material	0	0.999
3	Housing	0.17	0.681
4	Floor Type	4.927	0.085
5	Milking Method	7.418	0.006
6	Milking Routine	8.03	0.005
7	Washing of Udder	0.865	0.352
8	Drying of Udder	0	0.989
9	Position of Udder	0.001	0.979
10	Condition of Udder	2.305	0.316
11	Presence of Ticks	0.023	0.879
12	Lesions (teat end lesions)	0.054	0.816
13	Dipping (pre \ post teat dipping)	0.265	0.607
14	Standing Position after Milking	0.059	0.808
15	Last (milking the mastitis cow last)	0.46	0.498
16	Use of Hormones	0.004	0.95
17	Use of Towel	0.205	0.651
18	History of Mastitis	0.042	0.838
19	Udder Hygiene Score	1.758	0.415
20	Feed Sharing	4.647	0.031
21	Manure Removal	2.606	0.106
22	Number of Attendees	4.013	0.045
23	Type of Herd	0.004	0.95
24	Age	0.031	0.861
25	Stage of Lactation	0.654	0.721

used this analysis. The results are shown in Table A [31] see appendix A. As location of farm, size of herd, housing and age are not significantly associated with the target variable these factors are discarded. There are total 3 factors named management system, milking routine and breed that are significantly associated with the target variable at 5 percent level of significance. While the rest of the factors are associated with the target variable at 1 percent level of significance.After chi square analysis between independent factors and dependent factors we end up with 24 factors. Now we checked the association between each independent factor to deal with multicollinarity issue. After chi square analysis the results are shown in Table 4.2-4.24. Table 4.2 represents the association of breed with other factors and the results reveal that breed is significantly associated with two factors named feed sharing and number of attendees at 5 percent level of significance and with two factors named milking method and milking routine at 1 percent level of significance

Table 4.3 represents the association of type of herd with other factors and the results reveal that type of herd is significantly associated with 17 factors at 1 percent level of significance. Table 4.4 represents the association of stage of lactation with other factors and the results reveal that stage of lactation is

Table 4.3: Association of Type of Herd with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Management System	23.621	0
2	Bedding Material	300.084	0
3	Housing	12.31	0
4	Floor Type	116.424	0
5	Milking Method	49.533	0
6	Milking Routine	0.371	0.542
7	Washing of Udder	139.701	0
8	Drying of Udder	129.639	0
9	Position of Udder	1.241	0.265
10	Condition of Udder	13.66	0.001
11	Presence of Ticks	1.225	0.268
12	Lesions (teat end lesions)	9.885	0.002
13	Dipping (pre \ post teat dipping)	116.895	0
14	Standing Position after Milking	68.535	0
15	Last (milking the mastitis cow last)	273.103	0
16	Use of Hormones	114.62	0
17	Use of Towel	238.766	0
18	History of Mastitis	3.458	0.063
19	Udder Hygiene Score	13.929	0.001
20	Feed Sharing	34.261	0
21	Manure Removal	127.493	0
22	Number of Attendees	0.033	0.856
23	Age	0.007	0.934
24	Stage of Lactation	0.869	0.647

significantly associated with history of mastitis at 5 percent level of significance. Table 4.5 represents the association of management system with other factors and the results reveal that management system is significantly associated with 11 factors named at 1 percent level of significance and with 1 factor named Last (milking the mastitis cow last) at 5 percent level of significance. Table 4.6 represents the association of bedding material with other factors and the results reveal that bedding material is significantly associated with 14 factors named at 1 percent level of significance and with 1 factor named Lesions (teat end lesions) at 5 percent level of significance. Table 4.7 represents the association of floor type with other factors and the results reveal that floor type is significantly associated with 13 factors named at 1 percent level of significance. Table 4.8 represents the association of milking method with other factors and the results reveal that milking method is significantly associated with 8 factors at 1 percent level of significance

Table 4.9 represents the association of milking routine with other factors and the results reveal that milking routine is significantly associated with 4 factors at 1 percent level of significance and with feed sharing at 5 percent level of significance. Table 4.10 represents the association of washing of udder with other factors and the results reveal that washing of udder is significantly associated with 12 factors at

Table 4.4: Association of Stage of Lactation with other factors

Sr.No	**Factors**	**Chi-Square Value**	**P-value**
1	Management System	0.298	0.861
2	Bedding Material	0.594	0.743
3	Housing	1.454	0.483
4	Floor Type	4.112	0.391
5	Milking Method	0.011	0.995
6	Milking Routine	0.677	0.713
7	Washing of Udder	0.06	0.97
8	Drying of Udder	1.236	0.539
9	Position of Udder	0.945	0.624
10	Condition of Udder	3.973	0.41
11	Presence of Ticks	1.184	0.553
12	Lesions (teat end lesions)	3.236	0.198
13	Dipping (pre \ post teat dipping)	0.438	0.803
14	Standing Position after Milking	0.366	0.833
15	Last (milking the mastitis cow last)	1.204	0.548
16	Use of Hormones	0.679	0.712
17	Use of Towel	1.471	0.479
18	History of Mastitis	6.563	0.038
19	Udder Hygiene Score	0.955	0.917
20	Feed Sharing	0.938	0.626
21	Manure Removal	0.09	0.956
22	Number of Attendees	0.532	0.767

Table 4.5: Association of Management System with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Bedding Material	49.255	0
2	Housing	7.016	0.008
3	Floor Type	102.029	0
4	Milking Method	1.704	0.192
5	Milking Routine	0.783	0.376
6	Washing of Udder	27.477	0
7	Drying of Udder	21.7743	0
8	Position of Udder	0.001	0.975
9	Condition of Udder	0.764	0.683
10	Presence of Ticks	0.773	0.379
11	Lesions (teat end lesions)	0.005	0.946
12	Dipping (pre \ post teat dipping)	19.771	0
13	Standing Position after Milking	27.477	0
14	Last (milking the mastitis cow last)	6.089	0.014
15	Use of Hormones	18.608	0
16	Use of Towel	63.574	0
17	History of Mastitis	0.006	0.936
18	Udder Hygiene Score	10.002	0.007
19	Feed Sharing	241.191	0
20	Manure Removal	0.159	0.69
21	Number of Attendees	56.727	0

Table 4.6: Association of Bedding Material with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Housing	7.654	0.006
2	Floor Type	151.547	0
3	Milking Method	41.219	0
4	Milking Routine	1.391	0.238
5	Washing of Udder	133.14	0
6	Drying of Udder	217.29	0
7	Position of Udder	2.319	0.128
8	Condition of Udder	14.488	0.001
9	Presence of Ticks	2.85	0.091
10	Lesions (teat end lesions)	5.963	0.015
11	Dipping (pre \ post teat dipping)	170.159	0
12	Standing Position after Milking	101.442	0
13	Last (milking the mastitis cow last)	250.028	0
14	Use of Hormones	100.025	0
15	Use of Towel	299.243	0
16	History of Mastitis	3.646	0.056
17	Udder Hygiene Score	20.196	0
18	Feed Sharing	34.955	0
19	Manure Removal	145.151	0
20	Number of Attendees	0.227	0.633

Table 4.7: Association of Floor Type with other factors

Sr.no	Factors	Chi-Square Value	P-value
1	Milking Method	32.212	0
2	Milking Routine	18.325	0
3	Washing of Udder	56.112	0
4	Drying of Udder	121.282	0
5	Position of Udder	5.77	0.056
6	Condition of Udder	18.282	0.001
7	Presence of Ticks	1.003	0.606
8	Lesions (teat end lesions)	4.556	0.103
9	Dipping (pre \ post teat dipping)	216.28	0
10	Standing Position after Milking	57.571	0
11	Last (milking the mastitis cow last)	57.261	0
12	Use of Hormones	54.166	0
13	Use of Towel	75.51	0
14	History of Mastitis	1.556	0.459
15	Udder Hygiene Score	8.659	0.07
16	Feed Sharing	84.472	0
17	Manure Removal	139.421	0
18	Number of Attendees	26.552	0

Table 4.8: Association of Milking Method with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Milking Routine	0.053	0.819
2	Washing of Udder	64.329	0
3	Drying of Udder	51.513	0
4	Position of Udder	0.282	0.596
5	Condition of Udder	0.67	0.715
6	Presence of Ticks	2.838	0.092
7	Lesions (teat end lesions)	0.554	0.457
8	Dipping (pre \ post teat dipping)	1.786	0.181
9	Standing Position after Milking	21.837	0
10	Last (milking the mastitis cow last)	94.592	0
11	Use of Hormones	22.342	0
12	Use of Towel	8.552	0.004
13	History of Mastitis	0.453	0.501
14	Udder Hygiene Score	3.793	0.15
15	Feed Sharing	0.018	0.894
16	Manure Removal	8.721	0.003
17	Number of Attendees	27.8	0

Table 4.9: Association of Milking Routine with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Washing of Udder	0.051	0.822
2	Drying of Udder	0.188	0.665
3	Position of Udder	0.413	0.521
4	Condition of Udder	0.356	0.837
5	Presence of Ticks	1.628	0.202
6	Lesions (teat end lesions)	1.884	0.17
7	Dipping (pre \ post teat dipping)	29.484	0
8	Standing Position after Milking	1.767	0.184
9	Last (milking the mastitis cow last)	1.183	0.277
10	Use of Hormones	3.54	0
11	Use of Towel	18.864	0
12	History of Mastitis	0.057	0.811
13	Udder Hygiene Score	4.465	0.107
14	Feed Sharing	6.057	0.014
15	Manure Removal	0.012	0.912
16	Number of Attendees	22.647	0

Table 4.10: Association of Washing of Udder with other factors

Sr.No	**Factors**	**Chi-Square Value**	**P-value**
1	Drying of Udder	149.364	0
2	Position of Udder	4.601	0.032
3	Condition of Udder	21.406	0
4	Presence of Ticks	1.448	0.229
5	Lesions (teat end lesions)	13.457	0
6	Dipping (pre \ post teat dipping)	167.512	0
7	Standing Position after Milking	234.604	0
8	Last (milking the mastitis cow last)	138.114	0
9	Use of Hormones	208.565	0
10	Use of Towel	64.766	0
11	History of Mastitis	3.608	0.057
12	Udder Hygiene Score	31.483	0
13	Feed Sharing	18.42	0
14	Manure Removal	219.425	0
15	Number of Attendees	46.547	0

Table 4.11: Association of Drying of Udder with other factors

Sr.No	**Factors**	**Chi-Square Value**	**P-value**
1	Position of Udder	1.357	0.244
2	Condition of Udder	8.705	0.013
3	Presence of Ticks	0.754	0.385
4	Lesions (teat end lesions)	4.708	0.03
5	Dipping (pre \ post teat dipping)	191.933	0
6	Standing Position after Milking	154.382	0
7	Last (milking the mastitis cow last)	164.344	0
8	Use of Hormones	123.802	0
9	Use of Towel	129.595	0
10	History of Mastitis	1.109	0.292
11	Udder Hygiene Score	12.536	0.002
12	Feed Sharing	5.679	0.017
13	Manure Removal	133.327	0
14	Number of Attendees	23.27	0

1 percent level of significance and with history of mastitis at 5 percent level of significance Table 4.11 represents the association of drying of udder with other factors and the results reveal that drying of udder is significantly associated with 6 factors at 1 percent level of significance and with 3 factors named condition of udder, lesions (teat end lesions) and feed sharing at 5 percent level of significance.

Table 4.12 represents the association of position of udder with other factors and the results reveal that position of udder is significantly associated with 4 factors at 1 percent level of significance and with dipping (pre post teat dipping) at 5 percent level of significance. Table 4.13 represents the association of condition of udder with other factors and the results reveal that condition of udder is significantly associated with 10 factors at 1 percent level of significance. Table 4.14 represents the association of presence of ticks with other factors and the results reveal that presence of ticks is significantly associated

Table 4.12: Association of Position of Udder with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Condition of Udder	84.002	0
2	Presence of Ticks	21.019	0
3	Lesions (teat end lesions)	37.95	0
4	Dipping (pre \ post teat dipping)	5.308	0.021
5	Standing Position after Milking	3.574	0.059
6	Last (milking the mastitis cow last)	3.192	0.074
7	Use of Hormones	2.821	0.093
8	Use of Towel	2.444	0.118
9	History of Mastitis	15.353	0
10	Udder Hygiene Score	2.483	0.289
11	Feed Sharing	0.219	0.64
12	Manure Removal	2.509	0.113
13	Number of Attendees	0.005	0.946

Table 4.13: Association of Condition of Udder with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Presence of Ticks	27.513	0
2	Lesions (teat end lesions)	151.637	0
3	Dipping (pre \ post teat dipping)	17.098	0
4	Standing Position after Milking	18.895	0
5	Last (milking the mastitis cow last)	18.959	0
6	Use of Hormones	21.432	0
7	Use of Towel	18.391	0
8	History of Mastitis	25.098	0
9	Udder Hygiene Score	17.139	0.002
10	Feed Sharing	0.276	0.871
11	Manure Removal	10.86	0.004
12	Number of Attendees	3.114	0.211

Table 4.14: Association of Presence of Ticks with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Lesions (teat end lesions)	26.883	0
2	Dipping (pre \ post teat dipping)	1.429	0.232
3	Standing Position after Milking	2.312	0.128
4	Last (milking the mastitis cow last)	4.677	0.031
5	Use of Hormones	4.782	0.029
6	Use of Towel	0.936	0.333
7	History of Mastitis	0.278	0.598
8	Udder Hygiene Score	1.173	0.556
9	Feed Sharing	0.704	0.402
10	Manure Removal	0.432	0.511
11	Number of Attendees	0.13	0.718

Table 4.15: Association of Lesions (teat end lesions) with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Dipping (pre \ post teat dipping)	10.251	0.001
2	Standing Position after Milking	13.457	0
3	Last (milking the mastitis cow last)	6.969	0.008
4	Use of Hormones	15.858	0
5	Use of Towel	3.968	0.046
6	History of Mastitis	15.326	0
7	Udder Hygiene Score	6.578	0.037
8	Feed Sharing	0.206	0.65
9	Manure Removal	9.786	0.002
10	Number of Attendees	1.678	0.195

with lesions (teat end lesions) at 1 percent level of significance and with two factors named last (milking the mastitis cow last) and use of hormones at 5 percent level of significance. Table 4.15 represents the association of lesions (teat end lesions) with other factors and the results reveal that lesions (teat end lesions) is significantly associated 6 factors at 1 percent level of significance and with two factors named use of towel and udder hygiene score at 5 percent level of significance. Table 4.16 represents the association of dipping (pre post teat dipping) with other factors and the results reveal that dipping (pre post teat dipping) is significantly associated 8 factors at 1 percent level of significance and with feed sharing at 5 percent level of significance. Table 4.17 represents the association of standing position after milking

Table 4.16: Association of Dipping (pre post teat dipping) with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Standing Position after Milking	111.55	0
2	Last (milking the mastitis cow last)	146.8	0
3	Use of Hormones	105.669	0
4	Use of Towel	92.842	0
5	History of Mastitis	8.452	0.004
6	Udder Hygiene Score	18.001	0
7	Feed Sharing	5.365	0.021
8	Manure Removal	129	0
9	Number of Attendees	12.856	0

Table 4.17: Association of Standing Position after Milking with other factors

Sr.no	Factors	Chi-Square Value	P-value
1	Last (milking the mastitis cow last)	91.578	0
2	Use of Hormones	301.694	0
3	Use of Towel	43.944	0
4	History of Mastitis	6.496	0.011
5	Udder Hygiene Score	55.45	0
6	Feed Sharing	22.236	0
7	Manure Removal	106.116	0
8	Number of Attendees	13.585	0

Table 4.18: Association of Last (milking the mastitis cow last) with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Use of Hormones	140.801	0
2	Use of Towel	189.169	0
3	History of Mastitis	5.052	0.025
4	Udder Hygiene Score	19.55	0
5	Feed Sharing	0.354	0.552
6	Manure Removal	103.265	0
7	Number of Attendees	8.691	0.003

with other factors and the results reveal that standing position after milking is significantly associated 7 factors at 1 percent level of significance and with history of mastitis at 5 percent level of significance. Table 4.18 represents the association of last (milking the mastitis cow last) with other factors and the results reveal that last (milking the mastitis cow last) is significantly associated 5 factors at 1 percent level of significance and with history of mastitis at 5 percent level of significance. Table 4.19 represents the association of use of hormones with other factors and the results reveal that use of hormones is significantly associated 4 factors at 1 percent level of significance and with 2 factors named history of mastitis and no of attendees at 5 percent level of significance. Table 4.20 represents the association of use of towel with other factors and the results reveal that use of towel is significantly associated 3 factors at 1 percent level of significance and with factor no of attendees at 5 percent level of significance. Table 4.21 represents the association of history of mastitis with other factors and the results reveal that history of mastitis is significantly associated with factor, udder hygiene score at 1 percent level of significance.

Table 4.19: Association of Use of Hormones with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Use of Towel	42.95	0
2	History of Mastitis	4.73	0.03
3	Udder Hygiene Score	35.165	0
4	Feed Sharing	45.785	0
5	Manure Removal	78.148	0
6	Number of Attendees	4.646	0.031

Table 4.20: Association of Use of Towel with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	History of Mastitis	2.838	0.092
2	Udder Hygiene Score	22.113	0
3	Feed Sharing	72.321	0
4	Manure Removal	73.769	0
5	Number of Attendees	5.507	0.019

Table 4.21: Association of History of Mastitis with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Udder Hygiene Score	11.228	0.004
2	Feed Sharing	0.259	0.611
3	Manure Removal	0.159	0.69
4	Number of Attendees	0.013	0.908

Table 4.22 represents the association of use of udder hygiene score with other factors and the results reveal that udder hygiene score is significantly associated with factor named manure removal at 1 percent level of significance and with factor named no of attendees at 5 percent level of significance. Table 4.23 represents the association of feed sharing with other factors and the results reveal that feed sharing is significantly associated factor named number of attendees at 1 percent level of significance. Table 4.24 represents the association of manure removal with other factors and the results reveal that manure removal is significantly associated factor named number of attendees at 1 percent level of significance.

4.2.2 Selection of highly associated pairs after chi-square analysis

From the above analysis we selected two highly associated independent factors from each table that formed a pair on basis of chi-square analysis. In the pair the factor that is less associated with the target variable is discarded. For example our first pair is breed and milking routine, these are highly associated with each other. As the association of milking routine is more with target variable as compared to breed, so breed is discarded. Same steps are followed for other pairs as shown in Table 4.25. After these steps we end up with 10 factors. These factors are shown in Table 4.26. After factor selection on basis of chi-square analysis, we selected the factors based on assessment of laboratory professionals from each pair. These selected factors are shown in Table: 4.27. At this point we end with four subsets of factors the first

Table 4.22: Association of Udder Hygiene Score with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Feed Sharing	1.462	0.482
2	Manure Removal	16.361	0
3	Number of Attendees	6.678	0.035

Table 4.23: Association of Feed Sharing with other factors

Sr.no	Factors	Chi-Square Value	P-value
1	Manure Removal	3.384	0.066
2	Number of Attendees	45.339	0

Table 4.24: Association of Manure Removal with other factors

Sr.No	Factors	Chi-Square Value	P-value
1	Number of Attendees	31.724	0

Table 4.25: Highly associated pairs after chi-square analysis

Sr.No	Factors	Chi Square	P Value	Discarded Factors
1	Breed	4.67	0.03	Breed
	Milking Routine	5.45	0.02	
2	Herd Type	26.58	0	Herd Type
	Bedding Material	28	0	
3	Lactation Stage	11.95	0.003	History Of Mastitis
	History Of Mastitis	10.6	0	
4	Management System	5.17	0.02	Management System
	Feed Sharing	7.16	0.007	
5	Bedding Material	28	0	Use Of Towel
	Use Of Towel	15.7	0	
6	Floor Type	14	0.001	Floor Type
	Pre Post Teat Dipping	38	0	
7	Milking Method	9.5	0	Milking Method
	Milking Mastitis Cow Last	38	0	
8	Milking Routine	5.45	0.02	Milking Routine
	Pre Post Teat Dipping	38	0	
9	Washing Of Udder	46	0	Standing Position After Milking
	Standing Position After Milking	46	0	
10	Drying Of Udder	11	0	Drying Of Udder
	Pre Post Teat Dipping	38	0	
11	Udder Position	20.52	0	Udder Position
	Udder Condition	97.9	0	
12	Udder Condition	97.9	0	Teat End Lesions
	Teat End Lesions	69	0	
13	Presence Of Ticks	26	0	Teat End Lesions
	Teat End Lesions	69	0	
14	Pre Post Teat Dipping	38	0	Pre Post Teat Dipping
	Milking Mastitis Cow Last	38	0	
15	Standing Position After Milking	46.02	0	Standing Position After Milking
	Use Of Hormones	50.06	0	
16	Milking Mastitis Cow Last	38	0	Use Of Towel
	Use Of Towel	15.7	0	
17	Use Of Hormones	50.06	0	Manure Removal
	Manure Removal	12.12	0	
18	History Of Mastitis	10.6	0	History Of Mastitis
	Udder Hygiene Score	44.05	0	
19	Udder Hygiene Score	44.05	0	Manure Removal
	Manure Removal	12.12	0	
20	Feed Sharing	7.16	0.007	No Of People Attending
	No Of People Attending	7.14	0	

Table 4.26: Shortlisted factors based on Chi-square analysis

Sr.No	**Factors**
1	Bedding Material
2	Washing of Udder
3	Condition of Udder
4	Dipping (pre \ post teat dipping)
5	Last (milking the mastitis cow last)
6	Use of Hormones
7	Udder Hygiene Score
8	Feed Sharing
9	Stage of Lactation
10	Presence of Ticks

Table 4.27: Assessment based on laboratory professionals

Sr.No	**Factors**
1	Milking Routine
2	Bedding Material
3	History Of Mastitis
4	Dipping (pre \ post teat dipping)
5	Last (milking the mastitis cow last)
6	Management System
7	Milking Method
8	Washing Of Udder
9	Condition of Udder
10	Lesions (teat end lesions)
11	Use Of Hormones
12	Udder hygiene Score
13	Manure Removal
14	Feed Sharing

subset consists of 28 factors without any exclusion, second subset of factors consists of 10 factors based on chi-square analysis third subset consists of 14 factors based on assessment of laboratory professionals, forth subset is the union of statistically selected factors and subject knowledge represented in Table 4.28.

Table 4.28: Short-listed factors based on union of statistically selected features and subject knowledge

Sr.No	**Factors**
1	Feed Sharing
2	Washing of Udder
3	Condition of Udder
4	Dipping (pre \ post teat dipping)
5	Last (milking the mastitis cow last)
6	Use of Hormones
7	Udder hygiene Score

Table 4.29: Performance Evaluation of ML Models

Sr. No.	Factors	Model	Accuracy	Specificity	Sensitivity	Precision	Cross Validation
1	28	KNN	80	88	37	62	85
2	28	SVM	87	92	70	72	84
3	28	ANN	86	91	67	71	90
4	10	KNN	82	88	55	59	85
5	10	SVM	85	91	60	71	84
6	10	ANN	85	91	65	67	89
7	14	KNN	84	90	59	66	85
8	14	SVM	87	92	66	79	85
9	14	ANN	87	91	70	67	90
10	7	KNN	84	90	61	65	86
11	7	SVM	86	91	65	73	85
12	7	ANN	86	91	70	67	90

Table 4.30: Performance Evaluation of ML Models based on subsets of factors

Sr. No.	Model	Subsets of Factors	Accuracy	Specificity	Sensitivity	Precision	Cross Validation
1	SVM	28	87	92	70	72	84
2	ANN	10	85	91	65	65	89
3	ANN	14	87	91	70	67	90
4	ANN	7	86	91	70	67	90

4.3 Model Development

After selection of attributes we implemented three different machine learning models on each subset of factors from Table 4.29. Our results revealed that SVM performed best based on 28 factors with 70 percent sensitivity while on the basis of 10, 14 and 7 factors ANN performed best as compared to the other models with 65 percent sensitivity with respect to 10 factors and 70 percent sensitivity with respect to 7 and 14 factors. (Table: 4.30) From the information present in the Table 4.30, we infer that performance of ANN with respect to 14 and 7 factors is almost same but we selected ANN based on 7 factors as the best model for classification of mastitis. The 7 factors which are deleted from 14 factors have no importance with respect to the classification of mastitis because there is no difference between the performance evaluation matrix with respect to 14 and 7 factors.

Chapter 5

Discussion and Conclusion

In this study, the significance of 28 categorical independent factors with respect to the prevalence of mastitis is investigated. There is a binary dependent variable with 0 being normal and 1 being diseased (mastitis). Presence of mastitis is screened in cows via machine learning techniques KNN, SVM and ANN. For this purpose, association of external factors with each other is determined by using chi-square analysis to find the main external factors associated with mastitis disease and then the association of shortlisted external factors with prevalence of mastitis is determined. After short-listing factors on basis of chi-square analysis, the factors based on assessment of laboratory professionals are selected. The combinations include all set of factors, then reduced subsets of 10 factors based on chi-square analysis. 14 factors based on assessment of laboratory professionals and 7 factors named feed sharing, washing of udder, condition of udder, dipping (pre post teat dipping), last (milking the mastitis cow last), use of hormones and udder hygiene score based on union of statistically selected features and subject knowledge. Three different machine learning models KNN, SVM and ANN on each subset of factors were implemented. SVM performed best based on 28 factors with 70 percent sensitivity while on the basis of 10, 14 and 7 factors ANN performed best as compared to the other models with 65 percent sensitivity with respect to 10 factors and 70 percent sensitivity with respect to 7 and 14 factors. For model validation, 10-fold cross validation scheme is used.

Table 5.1: List of 7 significant factors

Sr.No	**Factors**
1-	Feed Sharing
2-	Washing of Udder
3-	Condition of Udder
4-	Dipping (pre \ post teat dipping)
5-	Last (milking the mastitis cow last)
6-	Use of Hormones
7-	Udder hygiene Score

5.1 Conclusion

Following are the major conclusions of the study:

• In this study, we dealt with the multicollinarity issue and infer that there are total 7 factors that are significantly contributing towards the prevalence of mastitis. We also proposed a procedure consist of ANN based on 7 factors that classify mastitis best with 70 percent sensitivity

• Based on the provided details, the study recommends the use of ANN with following 7 factors (Table 5.1)to predict category of the target class.

5.2 Limitations

Following are the few limitations that exist in the data:

1- Dataset is constrained.

2- There exists class imbalance in the dataset.

5.3 Future Recommendation

Following are the few suggestions for future studies:

1- In light of the results of this study, we advise additional investigation by gathering more local data from other places and cities in Pakistan for model validation and stability. For ease of adoption by end users, the procedure might also be automated and turned into a web application. This would make it easier to create an all-encompassing plan to reduce local disease prevalence.

2- For future, we recommend that more dataset should be collected with respect to the diseased cows (mastitis) to deal with class imbalancing problem.

3- Further refinement needed with respect to sensitivity for model improvement.

Bibliography

[1] S. N. Garcia, B. I. Osburn, and J. S. Cullor, "A one health perspective on dairy production and dairy food safety," *One Health*, vol. 7, p. 100086, 2019.

[2] A. Rehman, L. Jingdong, A. A. Chandio, and I. Hussain, "Livestock production and population census in pakistan: Determining their relationship with agricultural gdp using econometric analysis," *Information Processing in Agriculture*, vol. 4, no. 2, pp. 168–177, 2017.

[3] W.-T. Yang, C.-Y. Ke, W.-T. Wu, R.-P. Lee, and Y.-H. Tseng, "Effective treatment of bovine mastitis with intramammary infusion of angelica dahurica and rheum officinale extracts," *Evidence-Based Complementary and Alternative Medicine*, vol. 2019, 2019.

[4] A. J. Bradley, "Bovine mastitis: an evolving disease," *The veterinary journal*, vol. 164, no. 2, pp. 116–128, 2002.

[5] M. Cobirka, V. Tancin, and P. Slama, "Epidemiology and classification of mastitis," *Animals*, vol. 10, no. 12, p. 2212, 2020.

[6] A. Hocine, R. Bouzid, H. Talhi, and D. Khelef, "An epidemiological study of bovine mastitis and associated risk factors in and around eltarf district, northeast algeria," *Veterinarska stanica*, vol. 52, no. 5, pp. 553–564, 2021.

[7] S. Taponen, E. Liski, A.-M. Heikkilä, and S. Pyörälä, "Factors associated with intramammary infection in dairy cows caused by coagulase-negative staphylococci, staphylococcus aureus, streptococcus uberis, streptococcus dysgalactiae, corynebacterium bovis, or escherichia coli," *Journal of dairy science*, vol. 100, no. 1, pp. 493–503, 2017.

[8] L. Fox, "Prevalence, incidence and risk factors of heifer mastitis," *Veterinary microbiology*, vol. 134, no. 1-2, pp. 82–88, 2009.

[9] M. Bilal, M. Iqbal, G. Muhammad, M. Avais, and M. Sajid, "Factors affecting the prevalence of clinical mastitis in buffaloes around faisalabad district (pakistan)," *Int. J. Agri. Biol*, vol. 6, pp. 185–187, 2004.

[10] M. Rahman, M. Bhuiyan, M. Kamal, and M. Shamsuddin, "Prevalence and risk factors of mastitis in dairy cows," *Bangladesh Veterinarian*, vol. 26, no. 2, pp. 54–60, 2009.

[11] L. Oliveira, C. Hulland, and P. Ruegg, "Characterization of clinical mastitis occurring in cows on 50 large dairy herds in wisconsin," *Journal of dairy science*, vol. 96, no. 12, pp. 7538–7549, 2013.

[12] B. J. Kitchen, "Bovine mastitis: milk compositional changes and related diagnostic tests," *Journal of Dairy Research*, vol. 48, no. 1, pp. 167–188, 1981.

[13] E. Royster and S. Wagner, "Treatment of mastitis in cattle," *Veterinary Clinics: Food Animal Practice*, vol. 31, no. 1, pp. 17–46, 2015.

[14] T. Bobbo, S. Biffani, C. Taccioli, M. Penasa, and M. Cassandro, "Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[15] R. M. Hyde, P. M. Down, A. J. Bradley, J. E. Breen, C. Hudson, K. A. Leach, and M. J. Green, "Automated prediction of mastitis infection patterns in dairy herds using machine learning," *Scientific reports*, vol. 10, no. 1, pp. 1–8, 2020.

[16] E. Karimuribo, J. Fitzpatrick, E. Swai, C. Bell, M. Bryant, N. Ogden, D. Kambarage, and N. French, "Prevalence of subclinical mastitis and associated risk factors in smallholder dairy cows in tanzania," *Veterinary Record*, vol. 163, no. 1, pp. 16–21, 2008.

[17] L. Fadul-Pacheco, H. Delgado, and V. E. Cabrera, "Exploring machine learning algorithms for early prediction of clinical mastitis," *International Dairy Journal*, vol. 119, p. 105051, 2021.

[18] A. Khan, A. Z. Durrani, A. Yousaf, J. A. Khan, M. Chaudhry, M. A. Khan, A. Khan, *et al.*, "Epidemiology of bovine sub-clinical mastitis in pothohar region, punjab, pakistan in 2018.," *Pakistan Journal of Zoology*, vol. 51, no. 5, 2019.

[19] R. Hussain, A. Khan, M. T. Javed, and F. Rizvi, "Possible risk factors associated with mastitis in indigenous cattle in punjab, pakistan," *Pak. Vet. J*, vol. 32, no. 4, pp. 605–608, 2012.

[20] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient knn classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016.

[21] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.

[22] D. Cheng, S. Zhang, Z. Deng, Y. Zhu, and M. Zong, "knn algorithm with data-driven k value," in *International Conference on Advanced Data Mining and Applications*, pp. 499–512, Springer, 2014.

[23] S. Yue, P. Li, and P. Hao, "Svm classification: Its contents and challenges," *Applied Mathematics-A Journal of Chinese Universities*, vol. 18, no. 3, pp. 332–342, 2003.

[24] T.-N. Do, V.-H. Nguyen, and F. Poulet, "Speed up svm algorithm for massive classification tasks," in *International conference on advanced data mining and applications*, pp. 147–157, Springer, 2008.

[25] R. Bala and D. Kumar, "Classification using ann: A review," *Int. J. Comput. Intell. Res*, vol. 13, no. 7, pp. 1811–1820, 2017.

[26] I. E. Livieris, "Improving the classification efficiency of an ann utilizing a new training methodology," in *Informatics*, vol. 6, p. 1, MDPI, 2018.

[27] J. Tan, J. Yang, S. Wu, G. Chen, and J. Zhao, "A critical look at the current train/test split in machine learning," *arXiv preprint arXiv:2106.04525*, 2021.

[28] J. J. Salazar, L. Garland, J. Ochoa, and M. J. Pyrcz, "Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy," *Journal of Petroleum Science and Engineering*, vol. 209, p. 109885, 2022.

[29] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation.," *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[30] K. Allix, T. F. Bissyandé, Q. Jérome, J. Klein, R. State, and Y. Le Traon, "Large-scale machine learning-based malware detection: confronting the" 10-fold cross validation" scheme with reality," in *Proceedings of the 4th ACM Conference on Data and Application Security and Privacy*, pp. 163–166, 2014.

[31] M. Zaib, "Mastitis in dairy cows, a case study.," *SINES*, 2022.

Chapter 6

Appendix

Appendix A

Sr.No	**Factors**	**Chi-Square Value**	**P-value**
1	Management System	5.17	0.023
2	Bedding Material	28.15	0
3	Housing	2.72	0.099
4	Floor Type	14.04	0.001
5	Milking Method	9.51	0.002
6	Milking Routine	5.45	0.02
7	Washing of Udder	46.02	0
8	Drying of Udder	11.05	0.001
9	Position of Udder	20.52	0
10	Condition of Udder	97.9	0
11	Presence of Ticks	26.38	0
12	Lesions (teat end lesions)	69.03	0
13	Dipping (pre \ post teat dipping)	38.53	0
14	Standing Position after Milking	46.02	0
15	Last (milking the mastitis cow last)	38.53	0
16	Use of Hormones	50.06	0
17	Use of Towel	15.77	0
18	History of Mastitis	10.6	0.001
19	Udder Hygiene Score	44.05	0
20	Feed Sharing	7.16	0.007
21	Manure Removal	12.12	0
22	Number of Attendees	7.14	0.008
23	Size of Herd	2.23	0.135
24	Type of Herd	26.58	0
25	Age	0.93	0.335
26	Breed	4.65	0.031
27	Stage of Lactation	11.95	0.003
28	Location of Farm	0.03	0.871