Advances in Artificial Intelligence and Machine Learning; Research 4 (1) 1991-2013 Received 27-12-2023; Accepted 04-03-2024; Published 11-03-2024

# Livestock Disease Data Management for E-Surveillance and Disease Mapping Using Cluster Analysis

**Mohammed Kemal Ahmed** mohammed.kemal@astu.edu.et
*Department of Software Engineering, Data Science & Big Data Lab,*
*Addis Ababa Science and Technology University,*
*Addis Ababa, Ethiopia*

**Durga Prasad Sharma** dp.shiv08@gmail.com
*AMUIT, MOEFDRE,*
*Under UNDP & MSRDC-MAISM under RTU Kota,*
*India*

**Hussein Seid Worku** hussien.seid@aastu.edu.et
*Department of Software Engineering, Data Science & Big Data Lab,*
*Addis Ababa Science and Technology University,*
*Addis Ababa, Ethiopia*

**Getinet Yilma** getinetyilma@gmail.com
*Department of Computer Science & Engineering,*
*School of Electrical Engineering & Computing,*
*Adama Science & Technology University,*
*Adama, Ethiopia*

**Achim Ibenthal**
*Faculty of Engineering and Health,*
*HAWK University of Applied Sciences and Arts,*
*Göttingen, Germany*

**Dharmveer Yadav** dharmveeryadav@sxcjpr.edu.in
*St.Xavier's College,*
*Nevta Mahapura road,*
*Jaipur, Rajasthan*

**Corresponding Author:** Mohammed Kemal Ahmed

## Abstract

Livestock is a crucial source of livelihood for Ethiopians. However, the sector's contribution to the economy is not as significant as expected. This is mainly due to the prevalence of livestock diseases caused by various pathogens posing a serious threat to local and national food security, reducing income, and impacting the livelihoods of livestock keepers. However, the sector has constraints on improved data management framework for enhanced livestock disease pattern analysis and e-surveillance.

**Citation:** Mohammed Kemal Ahmed et al. Livestock Disease Data Management for E-Surveillance and Disease Mapping Using Cluster Analysis. Advances in Artificial Intelligence and Machine Learning. 2024;4(1):114.

**Objective:** To control and manage livestock diseases and unlock the full potential of the livestock sector via improved data management, disease pattern analysis, and e-surveillance. This study investigates how Electronic Livestock Health Recording Systems (ELHRs) facilitate inclusive data management for uncovering disease patterns. The proposed ELHR framework investigated against various common software quality parameters such as completeness, inclusiveness, functionality, and consistency in livestock disease data management literature and evaluated against existing livestock data management frameworks. A dataset comprising 18,333 samples of livestock disease cases obtained from the ELHRs framework was also used for disease burden analysis.

**Results:** From the results, the proposed ELHR framework with its holistic focus said to bridge the software quality gaps in previous related specific focus frameworks. From the clustering results, the proposed ELHRs dataset improved disease burden mapping with a silhouette score of 98% compared to another framework which is 68%. Therefore the proposed ELHRs framework's information content manifests improved disease pattern analysis and e-surveillance performance.

**Conclusion:** ELHRs framework can assist in identifying trends and patterns in livestock disease data, ultimately leading to more effective disease diagnosing and management strategies, therefore the ELHRs framework has the potential to revolutionize livestock disease management, disease pattern analysis, and e-surveillance.

**Keywords:** Livestock disease data management, Software Quality, Disease mapping, E-surveillance, Cluster Analysis.

# 1. INTRODUCTION

## 1.1 Background

According to the Food and Agriculture Organization of the United Nations (FAO), livestock plays a vital role in rural households, providing food, income, and other important benefits, as well as contributing to human health and well-being [1]. Moreover livestock sub-sector in Ethiopia is not only vital for ensuring food security, but also provide traction power, generating rural income and employment at the household level, contributing to national economic growth through foreign exchange, and holds deep cultural significance [2]. Ethiopia has the largest livestock population in Africa, with over 200 million animals. According to a 2020 report by the Central Statistics Agency of the Federal Democratic Republic of Ethiopia, the country has 65 million cattle, 51 million goats, 49 million chickens, 40 million sheep, 11 million equines, and 8 million camels. Although most Ethiopians rely on livestock for their livelihood, the sector's contribution to household incomes and the national economy is lower than expected. Livestock diseases caused by bacteria, viruses, protozoa, and parasites cause significant losses in productivity [3].

According to numerous sources, livestock diseases pose a serious threat to local and national food security, reduce income, and impact the livelihoods of livestock keepers. For example, a study by [4], found that in Ethiopia's Oromia region, livestock diseases caused the deaths of nearly 50% of oxen, 50% of local cows, 81% of crossbred cows, 52% of local calves, and 60% of crossbred calves within 12 months. The most common causes of death were Contagious Caprine Pleuropneumonia (CCPP), Contagious Bovine Pleuropneumonia (CBPP), and Foot and Mouth Disease (FMD) [4]. While these losses are devastating for individual households, the total cost to the country is even greater. The author in [2], estimated that Ethiopia loses USD 27-45 million annually due to livestock diseases that affect the export market. Even in areas where there is a lack of knowledge about the epidemiology of diseases and seasonal variations, diseases can account for the majority of economic losses. The research studies in [2, 3], revealed that 70% of human diseases are food-borne diseases that are transmitted from animal to human. Hence, controlling animal diseases means indirectly minimizing the transmission of infectious diseases among communities.

Understanding disease epidemiological information makes it easier to deploy disease control measures. The scarcity of well-documented livestock health information in developing nations like Ethiopia poses a significant hurdle in obtaining comprehensive and timely data for making informed decisions based on data [5]. The proliferation of innovative technologies has led to a significant increase in digital livestock disease data in the past decade. The pace of digital innovation has been significant, transforming every sector of the economy, including animal production, health, and welfare. The incorporation of advanced digital technologies into animal healthcare holds immense promise for revolutionizing national veterinary services and enhancing their efficiency. However, the successful realization of these benefits relies on the establishment of robust electronic livestock health records (ELHRs). Effective electronic livestock disease management is crucial to quickly identify disease prevalence and, as a result, reduce the risk of disease and significant economic loss. Upon rigorous review of articles, it is clear that the recognition of the advantages of using machine learning and artificial intelligence, cloud computing, and IoT is expanding rapidly in health areas [6, 7].

### 1.2 Cluster Analysis

According to researchers in [8], cluster analysis is part of unsupervised machine learning, that segregates a set of data objects (or instances) into meaningful subclasses. Each subclass is a cluster, such that its objects are similar to one another, yet dissimilar to objects in other clusters.

There are various categories of cluster analysis methods. The selection of the method used for data analysis depends on the nature of the dataset, the computational complexity, and the specific need of the user, among others [2]. Recent trends indicate that hierarchical clustering draws the attention of scholars for grouping, predicting, and validating different disease categories [9]. Hierarchical clustering provides several advantages in that it does not require to specify the number of clusters. It is the most suitable unsupervised machine learning algorithm to cluster objects from small to large datasets. Clustering is performed either through agglomerative or divisive ways. Agglomerative hierarchical clustering starts by treating each data point as its own cluster. It then repeatedly merges the two most similar clusters together until all of the data points are in a single cluster [10]. The basic principle of agglomerative hierarchical clustering follows six steps, i) reading the dataset, ii) pre-processing, iii) checking the purity of the dataset, iv) computation of the proximity matrix, v) starting from each data point forming its own cluster, vi) merging of proximate clusters and update of the proximity matrix until convergence is reached.

### 1.3 Cluster Analysis for Livestock Disease Mapping

Statistical analysis of livestock disease clustering is crucial in various situations. These include investigating the cause of a disease, monitoring livestock diseases geographically, associating environmental factors like atmospheric temperature with diseases and species, and defining epidemiological investigations and interventions based on disease cluster analysis. Such analyses help in identifying patterns and similarities between diseases and their causal agents, which can aid in developing effective preventive measures. Therefore, it is imperative to conduct disease cluster analysis to mitigate the impact of diseases on both humans and animals.

Improving ELHRs improves the information quality, which in turn improves to uncover disease patterns. Therefore, ELHRs and cluster analysis for livestock disease burden mapping coined together can further improve disease surveillance by providing a rich set of information and giving insight to domain experts' respectively. Our work aims to address the following research questions, which will be conclusively answered upon completion of the study and subsequent analysis.1) How is the Electronic Livestock Health Recording (ELHR) system in veterinary practices commonly used? 2) To what extent livestock disease analysis can be improved via ELHRs? 3) What is the implication of ELHR on livestock disease burden mapping?

## 2. LITERATURE REVIEW

The study in [11], found that in Ethiopia lack of an organized electronic livestock disease data management system at the regional, zonal, and district levels makes it difficult to respond quickly to disease risks. It also proved that the lack of well-documented information in Ethiopia makes it difficult for domain researchers to obtain complete and accurate information about livestock diseases. However digital technologies have the potential to strengthen prevention, improve productivity, and overall animal health care. Additionally, electronic livestock disease data analytics are essential for stakeholders in the livestock sector to tailor their services to the needs of livestock owners and to turn data into actionable information. Electronic livestock disease management systems can help the livestock sector improve disease prediction, prevention, and control by providing timely, complete, and accurate information. Few researchers have addressed this need by proposing various electronic livestock health recording systems (ELHRs). However, these systems vary in scope, with some limited to reporting, others focused on specific species, and others involving a limited number of stakeholders. Additionally, not all ELHRs adequately reflect the core attributes of a data management framework. For instance author in [12], studied the performance of Vet Africa (a mobile-based application) in terms of completeness, timeliness, and accuracy for cattle disease. Similarly researchers in [13], built the Kenya Animal Bio Surveillance System(KABS) mobile app and integrated it into Kenya's domestic and wild animal surveillance systems by adopting existing data collection tools and targeting disease syndromes. Outbreak and Vaccination Reporting System (DOVAR) is a digital platform to manage and report disease outbreaks. Its main purpose is to conduct national passive surveillance through the collection of animal disease outbreaks and vaccination data on a monthly basis [14]. Nevertheless, this platform proves effective for managing livestock health data, it lacks the comprehensiveness and functionalities to address the entire range of livestock health-related information and beside it failed to provide timely livestock health-related information. However, almost all of these frameworks fail to provide full common data quality attributes to livestock disease research

According to [8, 15–17], the value of a data management framework can be evaluated based on various dimensions, including completeness, accuracy, timeliness, consistency, and accessibility. Similarly, work in [16], argued that Big Data Governance Framework should include timeliness, meaningfulness, reliability, completeness, sufficiency, and protection of the data. Thus, having well-designed livestock disease data management framework is the prerequisite for data analytics. Based on the information provided, it can be inferred that there is a significant level of interest in the implementation of Electronic Livestock Disease Recording (ELHRs) in Ethiopia's veterinary service provision. The adoption of ELHRs could potentially lead to improvements in livestock disease data management and enhance the overall efficiency of veterinary services. This is a positive development that could have far-reaching benefits for Ethiopia's livestock industry [11]. To make data-driven decisions and use livestock health data for various purposes in this digital age, we need a well-organized that provides up-to-date information, inclusive, and complete electronic livestock health recording systems with different functionalities.

Software quality attributes in general, data quality measurements in particular became essential for both managing and utilizing data for data-driven decisions. Studies in [8, 15–17], proposed various perspectives for dataset quality management, pre-processing, analytics, and visualization values that need to be considered during building standard frameworks. Specifically, the authors in [15, 16], revealed big data assessment dimensions such as timeliness, meaningfulness, completeness, sufficiency and etc. The work in [8], highlighted common software quality attributes that address dataset-related issues and their measurement dimensions. Working towards considering quality dimensions can further improve the quality of data-driven decisions in ELHRs, disease burden mapping, and e-surveillance tasks.

## 2.1 IT-Based Livestock Disease Data Management

The author in [10], reported that the proliferation of innovative technologies has led to a significant increase in digital livestock disease data in the past decade. The pace of digital innovation has been significant, transforming every livestock sector of the economy, including animal production, health, and welfare. The potential benefits of incorporating new digital technologies in animal health are convincing and likely lead to new models that make national veterinary services more efficient. This bears the potential to increase standards for animal welfare and health practice [10]. However, gaining the full potential advantages and anticipated results of the digital revolution is challenging in all sectors of general veterinary services. It requires in-depth field studies bridging optimized algorithms, infrastructure, and inclusive data management. In [18], the authors argued that the capability of easily capturing massive volumes of diagnostic data and health information has created the opportunity to identify patterns and risk factors not only for individual animals but also across herds, regions, and species. These insights have renovated the field of diagnostics into a tool of prevention, supporting animal health professionals to take immediate action with confidence. Here, the early notification of disease events or outbreaks determines effective prediction and containment strategies for livestock diseases.

The Animal Disease Notification and Investigation System (ADNIS), and the World Animal Health Information System (WAHIS) platforms from the World Organization for Animal Health (OIE) (World Organization For Animal Health, 2014) where members report relevant domestic animal or wildlife diseases, including zoonosis, identified or detected within their territory. The data is digital but manually entered, so there is a risk of errors and incompleteness. As revealed in [19], electronic data collection can effectively enhance disease surveillance at slaughterhouses in smallholder production systems. By utilizing electronic forms to gather and submit data, including information on animal demographics and condemnations, real-time data storage and access to a central database become possible. This facilitates informed decision-making and enables prompt responses to potential disease outbreaks. However, several studies have highlighted that comprehensive fact-based decision-making necessitates data collection from a diverse range of stakeholders. Relying solely on a limited number of sources can lead to biased information [20]. For instance, if we look close at the research conducted in [12], the system was effective for managing livestock disease data and conducting diagnosis, treatment, and reporting, but the study only evaluated its effectiveness for cattle diseases with limited stakeholders.

Similarly the research by [20], explores the establishment of an integrated animal health surveillance system in Tanzania, utilizing 13 data sources including livestock farmers, slaughter facilities, markets, commercial farms, veterinary shops, and electronic surveillance tools. The system aims to overcome limitations in individual data sources and enable quicker detection and response to potential outbreaks. However, the proposed system faces drawbacks such as limited coordination, manual data transmission, inconsistent data frequency, limited internet access, and resource constraints. The study also highlights the need for complementary data sources and the sustainability of technology use.

The research in [14], aimed to address the premature death of domesticated ruminants in Ethiopia by combining farmer survey data, living standards measurement data, and Disease Outbreak and Vaccination Reporting (DOVAR) dataset. The results showed that most herds experienced mortality and reproductive losses. However, the study found little agreement between the different datasets, limiting predictive scope. The study also indicated the need for more reliable and comprehensive data sources in future studies to address the issue of premature death in domesticated ruminants.

The work in [15], presented twelve data quality frameworks aiming on the definition, assessment, and improvement of data quality in data-driven decision-making for both researchers and industry practitioners. Data quality frameworks such as AIMQ (Assessment and Improvement of Information Quality), CDQ (Corporate Data Quality), HDQM (High-Dimensional Data Quality Management) have been assessed to address data quality issues [15]. However, data quality attributes are still hidden within the quality frameworks. A work in [16], proposed Big Data Governance Framework to include timeliness, meaningfulness, reliability, completeness, sufficiency, and protection of the data. The work emphasized collaboration between different government agencies and departments to share data and insights effectively. The authors addressed big data governance from a pension service perspective.

The work in [17], proposed a conceptual framework for data management in Business Intelligence (BI) with master data management, data sources, ETL process, data marts, and visualization components. The proposed framework addresses the limitations of existing frameworks by combining desirable features from existing frameworks. The proposed framework focuses on the storage, transformation, management, and BI dimensions, giving less emphasis on the quality of the data elements for BI and decision-making. A systematic review in [8], identified 165 measures to 97 different quality attributes. The quality attributes of maintainability, functionality, completeness, and many more are evaluated based on measures such as variability, reusability, commonality, and compositionality.

### 2.2 Clustering

Clustering is useful in several machine learning and data mining tasks, for example in medicine to identify different disease categories, and in biology to find genetic information [21]. Clustering for Pattern Recognition by [22, 23]. In [24], clustering is applied to identify the dairy farm outcome of the compound. Cluster analysis was also used by [25], to evaluate disease risk in pre-parturient dairy cattle. Researchers in [21], also applied cluster analysis to group breast cancer, Alzheimer's, and lung cancer from a large biological and medical dataset. Similarly authors in [26], applied expert knowledge and machine learning to classify livestock herd types, in their study, the authors used a self-organizing map (SOM), an unsupervised learning algorithm that reduces the dimensionality of high-dimensional input data while preserving its topological relationships. The study found that the SOM algorithm's graphical representation is interpretable by livestock experts, facilitating a participatory process of iteratively adding knowledge rules and visualizing the resulting structure. Researchers in [27], used hierarchical clustering analysis to predict 1-year mortality after starting hemodialysis. Hierarchical clustering provides a comprehensive classification process, allowing us to infer unknown shared features within each cluster. Application of Agglomerative Clustering is also used to analyze phylogenetically on Bacterium of Saliva by [28]. As a methodology characteristic extraction with N-mer frequency is applied to the collected DNA sequence data. Distance metrics were calculated and an agglomerative clustering algorithm was implemented following the extraction of characteristic features from the DNA sequences. To construct a phylogenetic tree, the authors applied the resulting clusters. They found that the three agglomerative linkage methods (single, complete, and average) produced mostly different distance levels in each dendrogram, but all of the methods produced the closest object. Based on the minimal Davies-Bouldin Index (DBI), the authors found that the average linkage method performed relatively better than the others. Study in [29], applied Agglomerative Hierarchical Clustering with Dynamic Time Warping to classify residential households' daily load curves based on their consumption patterns. The proposed approach used Dynamic Time Warping (DTW) to find the best match between two load curves, whereas Agglomerative Hierarchical Clustering (AHC) was used to get a better starting point for the cluster centers. The study also conducted a comparative analysis with other clustering techniques, such as K-means, K-medoids, and Gaussian Mixture Models (GMMs), using different distance metrics.

A thorough review has shown that hierarchical clustering approaches can be effectively employed in diverse business domains, including disease mapping.

## 3. RESEARCH METHODOLOGY

This section includes sources of the data, methods of data collection, techniques, and analysis tools and clustering algorithms with different distance metrics. Data for this research were collected through a nationwide census of animal health professionals. Following diagnosis and treatment of animals, participants submitted their findings to a designated Google Cloud Storage container specifically created for this research using their smart phone. A. Data Management Framework

After carefully reviewing research publications, we identified limitations in the current science of art [8, 12–17], and through discussions with key stakeholders, we explored their roles and proposed the following ELHRs, as shown in FIGURE 1. The diverse data values related to different stakeholders with detailed descriptions is also presented in TABLE 1.

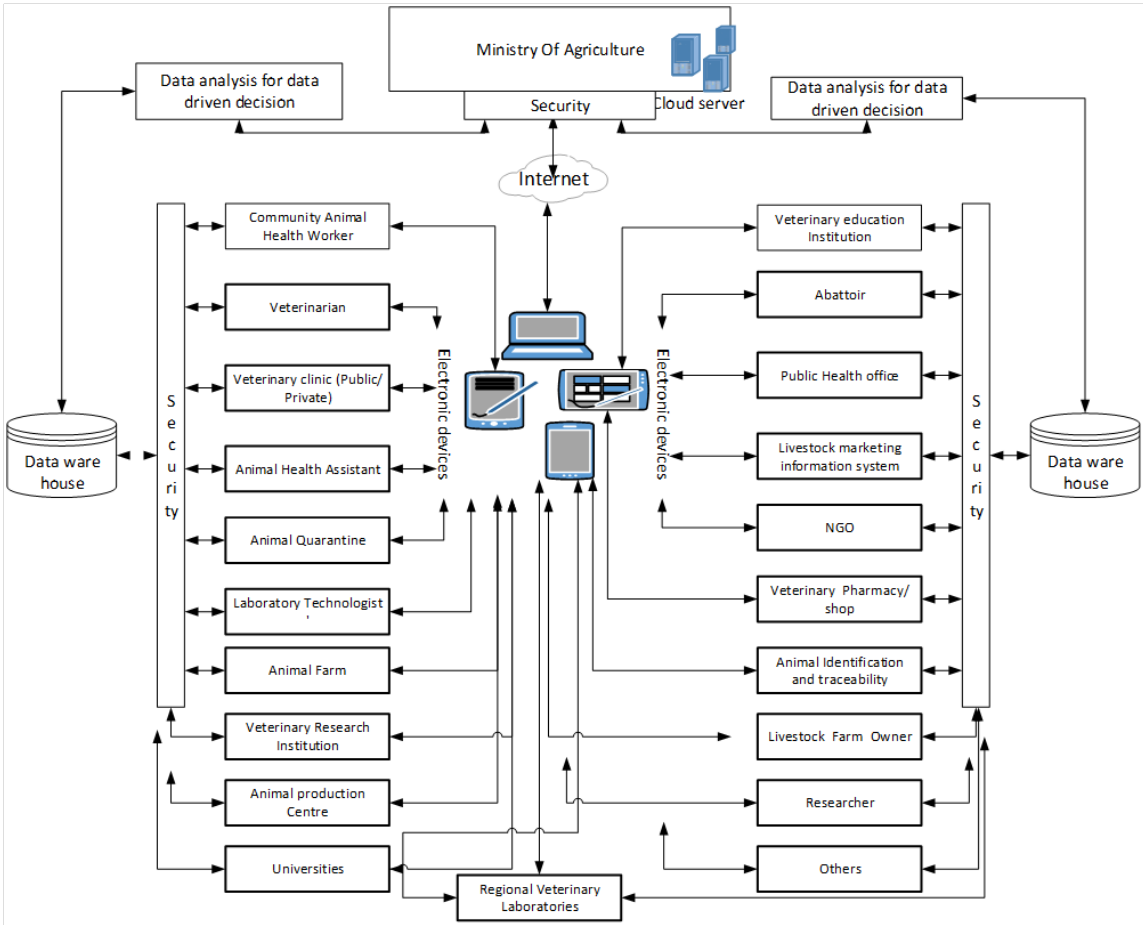

Ministry Of Agriculture
Cloud server
Security
Data analysis for data driven decision
Data analysis for data driven decision
Internet
Community Animal Health Worker
Veterinarian
Veterinary clinic (Public/ Private)
Animal Health Assistant
Animal Quarantine
Laboratory Technologist
Animal Farm
Veterinary Research Institution
Animal production Centre
Universities
Regional Veterinary Laboratories
Veterinary education Institution
Abattoir
Public Health office
Livestock marketing information system
NGO
Veterinary Pharmacy/ shop
Animal Identification and traceability
Livestock Farm Owner
Researcher
Others
Electronic devices
Electronic devices
Security
Security
Data ware house
Data ware house

Figure 1: Proposed Electronic Livestock Health Recording Systems (ELHRs)

### 3.1 Problem Formulation

1. How data quality (accuracy, completeness, consistency and timeliness) and data quantity (number of data points, number of features)
2. Identifying groups of livestock with similar disease patterns using cluster analysis.

In clustering, there are several commonly used metrics for characterizing distance similarity. As noted in [30], the Euclidean distance is the most intuitive distance metric because it corresponds to

Table 1: Data values, stakeholder and their description

| Data values [8, 15–17] | Name | Their description |
|---|---|---|
| Local disease data | Community Animal Health Worker(CAHW) [31] | Knowledgeable farmers, typically livestock owners, who are selected by their communities and receive training in basic animal health services to provide care at the village level. |
| Pathologies, prevalence, symptoms | Regional Veterinary Laboratory [32] | An organization provides diagnostic services to veterinarians and animal owners, including post-mortem examinations, clinical pathology, surgical biopsy interpretations, bacteriology, endocrinology, nutrition, parasitology, immunodiagnostics, and toxicology. |
| Pre and post diagnosis information | Animal Health Assistant [31] | Animal health technicians are essential members of the veterinary healthcare team, providing crucial support to veterinarians in the diagnosis and treatment of animals. |
| Accidental livestock health data | Animal Quarantine station [33] | A designated space for the temporary housing and examination of live animals. |
| Health incidents related to market | Livestock Marketing information System | A network of interconnected participants in the livestock industry, enabling the efficient and effective sale and purchase of livestock and their products. |
| Livestock production, medicine, policy, and support information | Livestock State Minister(LSM) [33] | A governmental entity within the Ministry of Agriculture of the Federal Republic of Ethiopia, tasked with spearheading and formulating policies that harness the potential of emerging technologies and proven methods to enhance the delivery of animal healthcare services. |
| Support, drugs, research and etc. | NGO [32] | A non-governmental organization is an entity that is typically established and operates autonomously from any governmental authority |
| Active health threats, symptoms, disease, training and preventive methods | Public Health Institute [32] | Is a scientifically-based organizations committed to advancing public health by coordinating functions and programs to prevent, detect, and respond to health threats. |
| Daily disease reporting | Private Veterinary clinic Owner | Private owned animal health facility |
| Active health threats, symptoms, disease, training and preventive methods | Veterinarian [5] | Is an expert who rehearses veterinary medication by treating sicknesses, issues, and wounds in non-human beings |
| Active health threats, symptoms, disease, training and preventive methods | Government Veterinary Clinic [5] | A government-owned animal health facility |
| Location, disease, heard information, productivity, and marketing | Livestock Owner (Farmers, Urban Dwellers, and Pastoralists) [31] | Those who rear and manage domesticated animals, from limited numbers to substantial herds, with the primary objective of deriving financial benefit |

the way we perceive distances in everyday life. The Euclidean distance D of two data cases (x1, y1) is defined as the square root of the sum of squared differences. It is a continuous metric that can be thought of in geometric terms as the “straight line” distance between two points.

In general, the formula for Euclidean distance between points (x1, y1) ∈ {X, Y} in the dataset is calculated as follows in an n-dimensional cluster space. The following are points on the 2-dimensional space,

$$\mathrm{X} = (\mathrm{x_1},\ \mathrm{x_2},\ \mathrm{x_3}, .., \mathrm{x_n})\ \mathrm{and Y} = (\mathrm{y_1}, \mathrm{y_2},\ \mathrm{y_3},\ ...., \mathrm{y_n}) \tag{1}$$

Hence, the Euclidean distance is

$$\mathrm{D_E^2} = (\mathrm{x_1} - \mathrm{y_1})^2 + (\mathrm{x_2} - \mathrm{y_2})^2 + .. + (\mathrm{x_n} - \mathrm{y_n})^2 \tag{2}$$

$$\mathrm{D_E}\,(\mathrm{X},\ \mathrm{Y}) = \sqrt{(\mathrm{x_1} - \mathrm{y_1})^2 + (\mathrm{x_2} - \mathrm{y_2})^2 + .. + (\mathrm{x_n} - \mathrm{y_n})^2} \tag{3}$$

Among other distance metrics, Manhattan distance uses the sum of the absolute differences between the coordinates of a pair of objects (X, Y). It is relatively efficient and easily understandable straightforward forward and gives the best results when the data set has a high dimension mathematically represented as.

$$D_M\,(X, Y) = \sum_{i=1}^{n} |x_i y_i| \tag{4}$$

Cosine distance, as defined in [32], is a measure of the angle between two vectors. It is calculated as follows:

$$cosinsimilarity = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5}$$

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a data set into k groups without a predefined target class. It defines the total within-cluster variation as the sum of squared distances between the corresponding centroid, mathematically computed as

$$\mathrm{W}\,(\mathrm{C_k}) = \sum_{\mathrm{x_i} \in \mathrm{C_k}} (\mathrm{x_i} - \mu_\mathrm{k})^2 \tag{6}$$

Where, $\mathrm{x_i}$ is a data point belonging to the cluster $\mathrm{c_k}$, μk is the mean value of the points assigned to the cluster C. Each observation ($\mathrm{x_i}$) is assigned to the given cluster such that the Sum of Squares (SS) distance of the observation to their assigned cluster centres $\mu_\mathrm{k}$ is a minimum. We can define the total within-cluster variation as follows:

The total within a cluster is represented by

$$\sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{W}\,(\mathrm{C_k}) = \sum_{\mathrm{k=1}}^{\mathrm{kK}} \sum_{\mathrm{x_i} \in \mathrm{C_k}} (\mathrm{x_i} - \mu_\mathrm{k})\,2 \tag{7}$$

The total within-cluster sum of squares (WSS) is a measure of the compactness of a clustering. It is calculated by summing the squared distances between each data point and the mean of its cluster. The WSS should be as small as possible, which means that the data points within each cluster should be as close together as possible. Hierarchical clustering can be either agglomerative (bottom-up), where smaller clusters are merged into larger clusters, or divisive (top-down), where larger clusters are split into smaller clusters it uses different linkage methods to measure the distance between two sub-clusters of data points. The most common linkage types are: Single, Complete and Average [34].

Single linkage clustering: Single linkage clustering is a hierarchical clustering method that merges clusters based on the shortest distance between any two data points in the clusters. This means that the two clusters with the closest data points will be merged first.

Assuming that disease of **“s”** represents cluster 1 and disease of **“t”** represents cluster 2 in the FIGURE 2, single linkage can be computed as follows:

$$d_{C_s,C_t} = argmin\ sin\,(st)\,(min\ D\,(x,y)\ \ where\ \ \ x \in C_s\ ,\ y \in C_t) \tag{8}$$

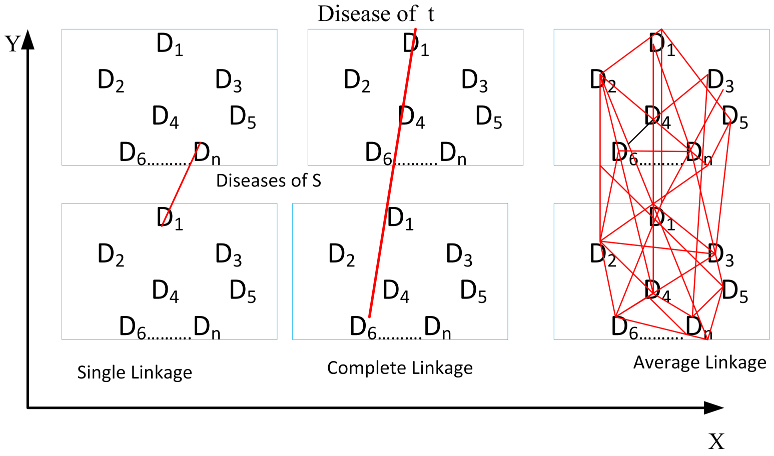


Figure 2: Different linkage methods in hierarchical clustering for mapping livestock diseases

Complete linkage clustering is a hierarchical clustering method that finds the maximum possible distance between any two points in different clusters.

$$d_{C_s,C_t} = argmin\ sin\,(s,t)\,(max\ D\,(x,y)\ \ where\ \ \ x \in C_s\ ,\ y \in C_t) \tag{9}$$

Average linkage: Find all possible pairwise distances for points belonging to two different clusters and then calculate the average.

$$d_{C_s,C_t} = argmin\ sin\,(s,t)\,\frac{1}{|C?|}\,\frac{1}{|C?|}(\sum x \in C?\ \sum y \in C?\ D\,(x,y)) \tag{10}$$

### 3.2 Sources and Descriptions of the Data

This project uses cluster analysis to map the distribution of livestock diseases in Ethiopia. The data for the analysis was collected from different regions, zones, and districts of the country using the ELHRs platform. In addition, for the purpose of comparison, we extracted data from DOVAR [14], and OIE frameworks. As shown in TABLE 2, the dataset characteristics is described in detail.

Table 2: Data source description

| Item | Source/Selection/Quantification | | |
|---|---|---|---|
| Data source | ELHRs | DOVAR [14] | OIE[1] |
| Selected Sources | Multiple Sectors | Ministry of Agriculture | FAO Livestock Sector |
| Selected Species | avian, bovine, caprine, equine, ovine | avian, bovine, caprine, equine, ovine | avian, bovine, caprine, equine, ovine |
| Selected Diseases | African Horse Sickness (AHS), Black Quarter/leg (BQ), Contagious Bovine Pleuropneumonia (CBPP), Hemorrhagic Septicemia (HS), Lumpy Skin Disease (LSD), Newcastle Disease (NCD), Sheep and Goat Pox (SGPX), Epizootic lymphangitis, GI parasites, Trichostrongylosis, Lungworm /Verminous Pneumonia | African Horse Sickness (AHS), Black Quarter/leg (BQ), Contagious Bovine Pleuropneumonia (CBPP), Hemorrhagic Septicemia (HS), Lumpy Skin Disease (LSD), Newcastle Disease (NCD), Sheep and Goat Pox (SGPX) | African Horse Sickness (AHS), Black Quarter/leg (BQ), Cntagious Bovine Pleuropneumonia (CBPP), Hemorrhagic Septicemia (HS), Lumpy Skin Disease (LSD), Newcastle Disease (NCD), Sheep and Goat Pox (SGPX) |
| # of cases | 18333 | 240817 | 13009 |
| # animals at risk | 15570839 | NA | NA |

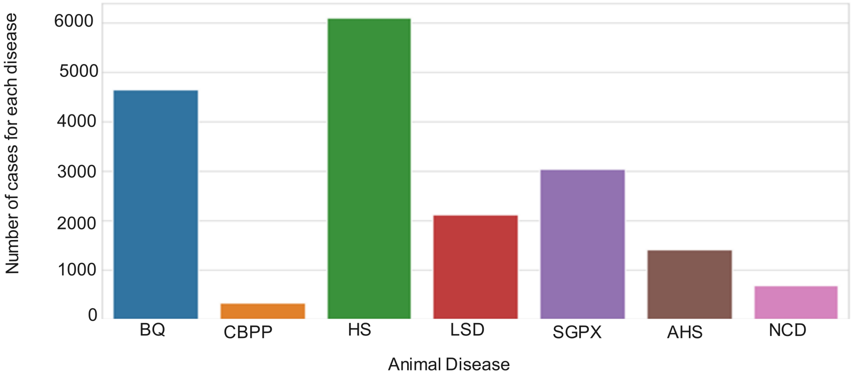


Figure 3: Number of cases of seven top diseases from ELHRs dataset.

FIGURE 3 indicates diseases that affect livestock severely in the form of disease out-breaks. As revealed in the Figure, the most frequent diseases that occurred were HS followed by BQ, SGPX, and LSD respectively. FIGURE 4 shows that the number of animals that died from BQ disease was high in Afar, Benishangul-Gumuz, Somalia, Gambela, and Tigray regions. The survey also found that BQ disease is most likely to occur in hot areas. BQ disease is a highly contagious and deadly disease that can affect livestock. It is important to be aware of the symptoms of BQ disease and to take steps to prevent it from spreading. The analysis results can help the sector design ways to minimize the risk of BQ disease. As revealed in FIGURE 5, the correlation analysis indicates that a positive

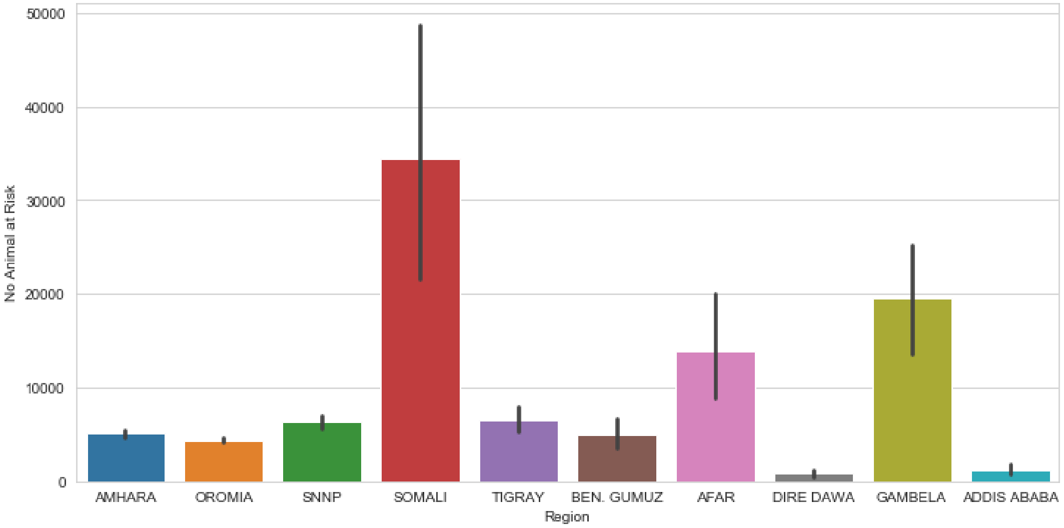


Figure 4: Number of animals which died at each region due to BQ disease.

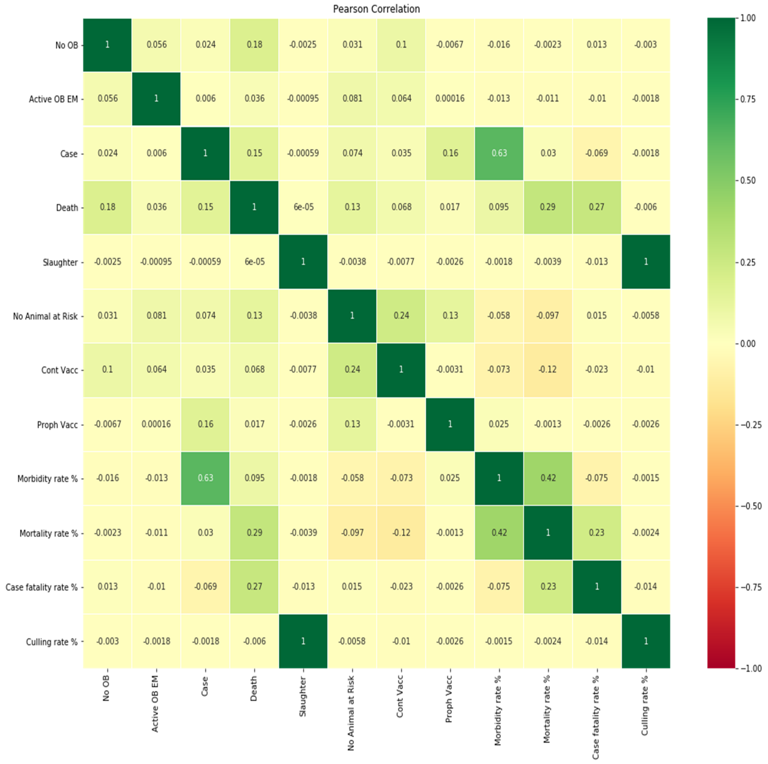


Figure 5: The correlation of the variables on Pearson correlation(ELHRs)

correlation between variables indicates a directional relationship between them, while a negative correlation indicates an inverse relationship and near to zero no significant relationship at all. For example, a correlation value of 1 between culling rate and slaughter indicates a perfect positive correlation, meaning that whenever there is slaughter, there is also culling of some or all internal organs due to various diseases. This information can be used by stakeholders in the relevant sector to investigate the underlying issues. Based on the observed negative correlation in the FIGURE 5, it appears that administering the control vaccine after the onset of symptoms might not effectively decrease morbidity (illness occurrence)

## 4. RESULTS

In this study, we extracted data from the ELHRs framework and preprocessed it using Python's sklearn preprocessing library. For clustering we used Python's sklearn, to compute distance metrics sklearn.metrics library was applied and matplotlib.pyplot to visualize the results.

Data pre-processing steps were taken to ensure the quality of the dataset, following the advice of a domain expert. These steps included cleaning, transforming, and normalizing the data, removing irrelevant values, and validating the data. The dataset was then loaded into a Jupiter Notebook to see which cluster algorithms were suited for mapping livestock disease burden mapping. K-means clustering with the elbow method, agglomerative hierarchical clustering with linkage methods, and different distance metrics were all computed.

The proposed ELHR framework addresses gaps in the current state of the art and explores potential stakeholders and their key roles. It is designed to support key stakeholders in managing livestock disease data for diagnosing and treating animals and to provide complete information to researchers, educators, and policymakers. The proposed ELHR framework was critically analyzed against previous livestock data management frameworks [12–14]. As shown in TABLE 3, the ELHRs was made to fill the gaps in existing data management frameworks providing information related to various species and disease, diverse stakeholder, and both data management and analytics roles. Therefore compared to previous data management frameworks, ELHRs have more software quality attributes addressing livestock disease data management and analysis needs.

Table 3: Data management framework data quality parameters

| Framework | Completeness | Inclusiveness | Functionalities | Consistency | Timeliness |
|---|---|---|---|---|---|
| DOVAR [14] | No | Yes | No | No | No |
| KABS [13] | No | No | No | Yes | Yes |
| Vet-Africa for Ethiopia [12] | Yes | No | No | Yes | Yes |
| OIE[2] | Yes | No | No | Yes | Yes |
| **ELHRs (Ours)** | Yes | Yes | Yes | Yes | Yes |

How do we select the right number of clusters? There are various methods that can be applied, but in this study we selected computing k-means clustering using different values of k also shown in FIGURE 6. The reason we used K means in order to get the optimal number of cluster using elbow method. Next, the Within Sum of Squares (WSS) was plotted according to the number of clusters. The location of a bend (knee) in the plot is commonly considered as a sign of the appropriate number of clusters. We can see a bend or elbow at k = 5 indicates that additional clusters beyond the fifth have little value, as the variance will not decrease significantly with further increases in k

[35]. Additionally, we applied silhouette analysis to evaluate the performance of different distance metrics because it is a method for measuring how well distance metrics group diseases together. As described in [36], Silhouette cluster analysis is a method for evaluating the quality of clustering algorithms. It measures the similarity of each point to its own cluster (cohesion) compared to other clusters (separation). Mathematically represented as follows

$$S\left(i\right)=\frac{\left(B\left(i\right)-A\left(i\right)\right)}{\max(\left(B\left(i\right),A\left(i\right)\right)}$$

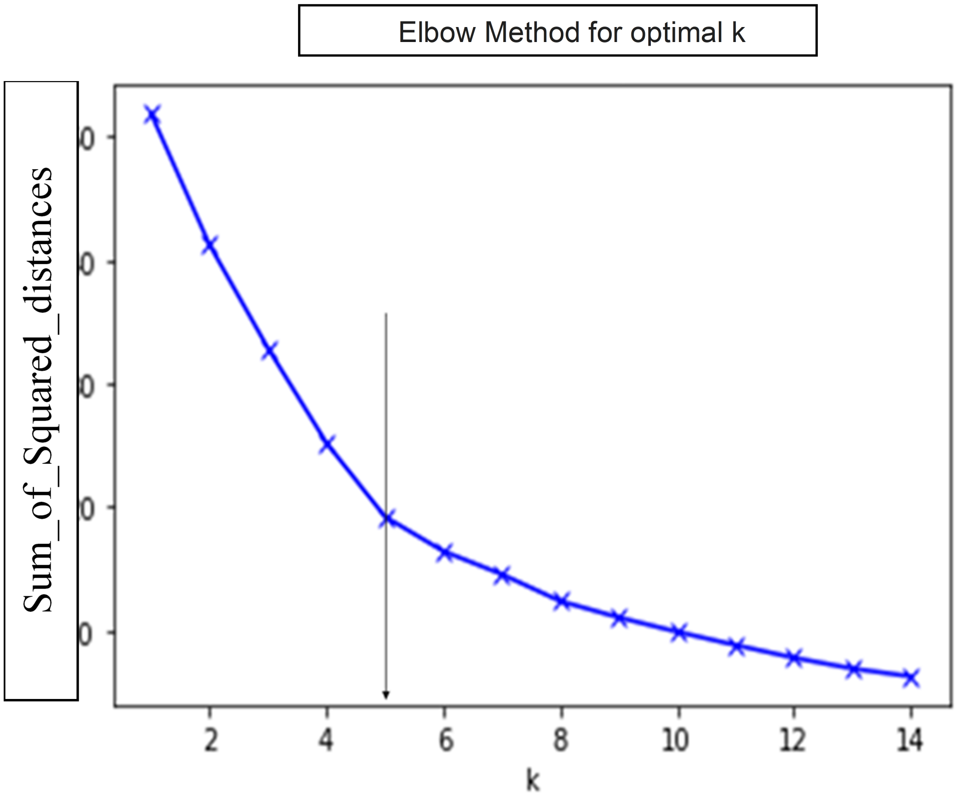


Figure 6: Clustering of disease using elbow method.

Where:

S (i) is the silhouette value of object (i)

B (i) is the average distance between object (i) and all other objects in its cluster

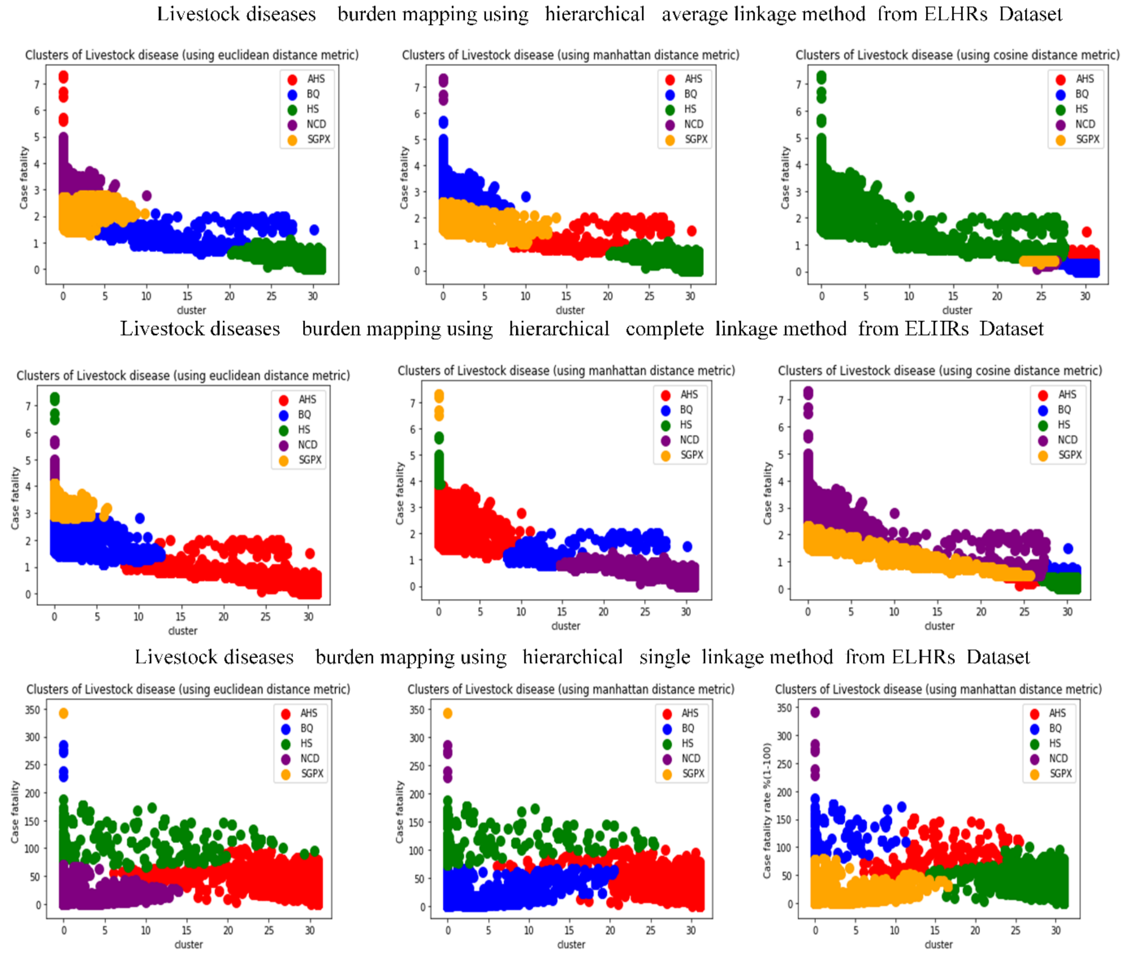
A (i) is the minimum average distance between object (i) and all objects in any other cluster

The value ranges between (−1 and 1) the higher the value is the better the clustering algorithms in grouping the objects [36].

The ELHRs dataset was evaluated using different linkage methods in different distance metrics as shown in TABLE 4. The silhouette values for Euclidean and Manhattan in single linkage were both 98%, while the silhouette score of cosine is 38%. The corresponding cluster results are indicated in FIGURE 7. It shows the results of cluster visualization of livestock disease burden using three linkage methods (average, complete, and single) and different distance metrics applied to the ELHRs data points.

Table 4: The silhouette score of different linkage methods in different distance metrics

| **Linkage Method** | **Distance Metrics** | | |
|---|---|---|---|
| | Euclidean | Manhattan | Cosine |
| Average | 80% | 78% | 60% |
| Complete | 80% | 75% | 60% |
| Single | 98% | 98% | 38% |


Livestock diseases burden mapping using hierarchical average linkage method from ELHRs Dataset

Livestock diseases burden mapping using hierarchical complete linkage method from ELIIRs Dataset

Livestock diseases burden mapping using hierarchical single linkage method from ELHRs Dataset

Figure 7: Clustering of livestock disease burden using different linkage methods in different distance metrics

To assess the effectiveness of the ELHRs framework for livestock disease burden mapping and provide insights for decision-makers, we evaluated a single linkage method using Euclidean distance metrics on different datasets. The silhouette score results are indicated in TABLE 5.

Table 5: The silhouette of the single linkage method on different dataset

| **Dataset** | **Silhouette Score** |
|---|---|
| DOVAR **[14]** | 67% |
| OIE | 68% |
| **ELHRs (Ours)** | 98% |

FIGURE 8 presents the results of clustering livestock disease burden using a single linkage method and Euclidean distance metrics applied to three datasets DOVAR [14], OIE and ELHRs(Ours). As revealed in the figure there are five clusters of diseases projected in different colures, each cluster represents a group of diseases with similar characteristics. The different colors help to visually distinguish between the clusters.

## 5. DISCUSSION

The proposed ELHRs framework as indicated in TABLE 3 bridges the software quality value gaps in existing frameworks. Whereas, existing frameworks leverage datasets with specific focus, ELHRs leverages towards holistic focus in which a data point improved information content via diverse feature distribution about a single data point which leads towards generalized disease pattern understanding and e-surveillance whereas, specific focus data point in DOVAR [14], KABS [13], Vet-Africa for Ethiopia [12], and OIE leverage limited feature distribution about a single data point leading to less generalization and more biased towards specific data points.

Current initiatives like Vet-Africa (focused on cattle) and DOVAR/OIE (providing monthly outbreak reports) offer valuable contributions. However, decision-makers, policymakers, and researchers require access to up-to-date data encompassing various species and diseases. In this regard, the ELHRs framework stands out by catering to this broader need for comprehensive and timely information.

The proposed ELHRs dataset for disease clustering and e-surveillance using single, average, and complete linkage clustering algorithms with different distance metrics. As shown in TABLE 4, all clustering algorithms provides promising silhouette score in livestock disease clustering. From the result however, single linkage clustering with Euclidian distance metric provided the best silhouette score 98% in disease clustering. From the result, higher silhouette score implies better clustering of disease burden mapping and hence improved control implication.

The clustering visualization shown from FIGURE 7 depicted projection of the ELHRs sample data points in two dimensional projection space. From the result it is clear each disease data points are projected in a separate space for single linkage method compared to complete and average linkage methods. From the clustering result it is possible to draw single linkage method's ability in livestock disease pattern analysis and e-surveillance improve disease control via giving insight to domain experts and policy makers.

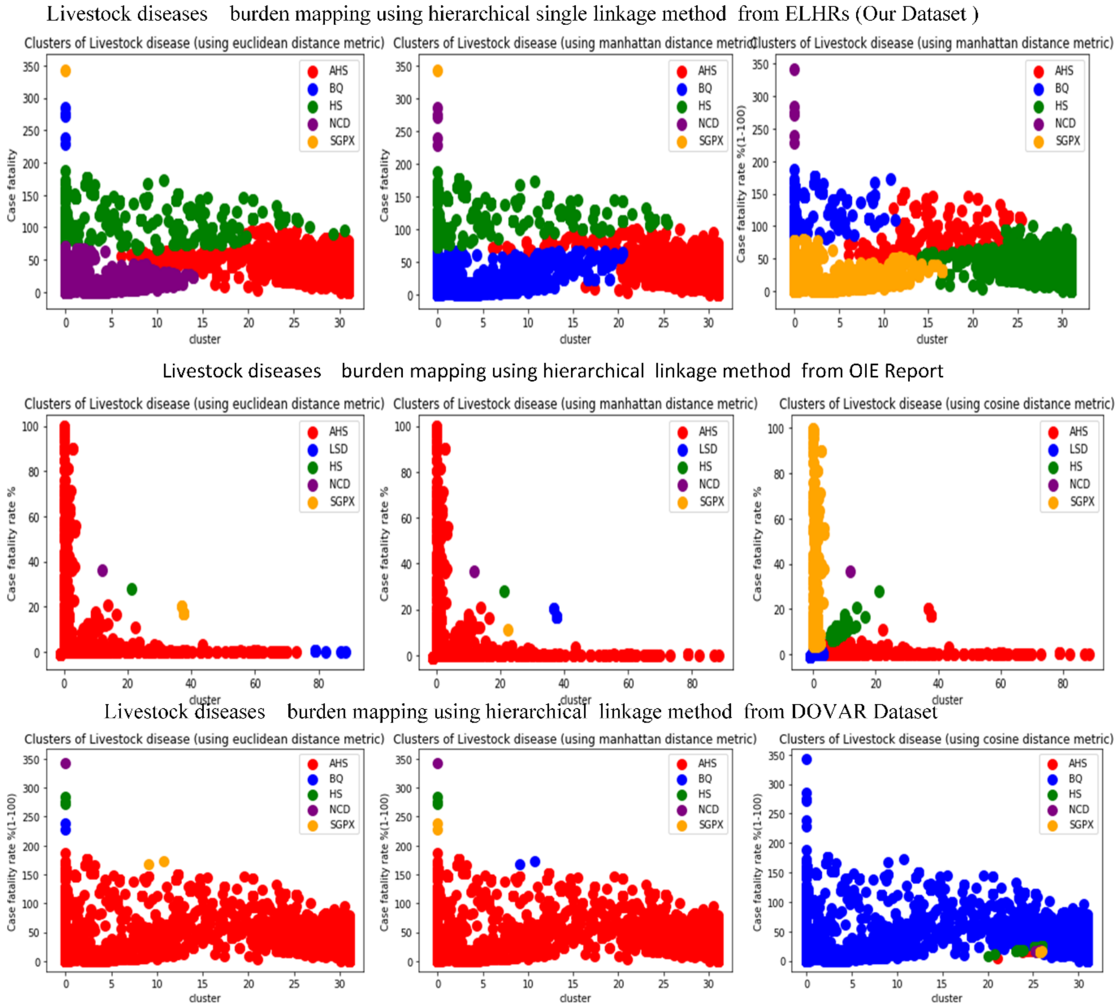


Figure 8: Clustering of livestock disease burden using single linkage methods in Euclidean distance metrics on different dataset

To evaluate, the performance of the proposed ELHRs framework against DOVAR [14], and OIE datasets, a single linkage clustering method with Euclidian distance metric was employed. From TABLE 5, the result shows, ELHRs dataset silhouette score surpassed both DOVAR [14], and OIE datasets by +30%. The improvement in ELHRs dataset suggests the dataset features, feature distribution and diversity is better than DOVAR and OIE dataset. Therefore, the proposed ELHRs framework software quality values have holistic focus which leads towards generalization in livestock disease clustering performance.

With the same token the two dimensional projection of the data points extracted from ELHRs, DOVAR [14], and OIE frameworks are depicted in FIGURE 8, using single linkage clustering method. When we see the clusters, the proposed ELHRs is able to project livestock disease data point in separate 2D embedding compared to the projection in DOVAR and OIE. The DOVAR and OIE projections show overlaps for some disease data points, similar disease data points are also projected in different 2D embedding spaces representing as an outlier, and DOVAR dataset feature distribution has more bias towards AHS and BQ livestock diseases. From clustering result, we can conclude the proposed ELHRs dataset feature spaces leverage improved information content for single linkage clustering method compared to data points extracted from DOVAR, and OIE.

To sum up The ELHRs framework is a holistic approach to disease clustering and e-surveillance, improving information content through diverse feature distribution. It uses single, average, and complete linkage clustering algorithms, with the single linkage clustering method providing the best silhouette score of 98%.

## 6. CONCLUSIONS

This work proposed ELHRS framework for livestock disease burden mapping and e-surveillance. The prime objective of the study was to show how an appropriate livestock disease data management system facilitates data analysis for disease mapping. The proposed method is designed to bridge the dataset quality, and management challenges in the state of the art livestock disease management frameworks such as DOVAR, OIE, vet-Africa, and KABS. The proposed ELHRs framework is therefore hypothesized to leverage holistic dataset values and management in the livestock disease burden mapping and e-surveillance tasks.

The effectiveness of the proposed data management framework is validated from software quality measurement literature. Furthermore, disease pattern analysis and e-surveillance performance is evaluated by training clustering algorithms using data extracted from the proposed ELHRs framework, and data from other related frameworks such as DOVAR and OIE.

From the results, the proposed ELHRs framework with its holistic focus said to bridge the software quality gaps in previous related specific focus frameworks. From the clustering results, the proposed ELHRs dataset improved disease burden mapping with a silhouette score 98% compared to OIE framework silhouette score which is 68%. The quality attributes of the proposed ELHRs dataset is also shown in 2D projection of data point feature distribution, which produced semantically separated disease clusters compared to OIE and DOVAR datasets feature projections which leads towards overlapping and clusters with outliers. Therefore the proposed ELHRs framework's holistic focus information content manifests improved disease pattern analysis and e-surveillance performance. ELHRs framework can assist in identifying trends and patterns in livestock disease data, ultimately leading to more effective disease diagnosing and management strategies, therefore the ELHRs framework has the potential to revolutionize livestock disease management, disease pattern analysis and e-surveillance.

### 6.1 Recommendation

While the study benefited from valuable data, it could be further enhanced by incorporating specific variables in the dataset from a wider range of stakeholders to extract deeper insights. Moving forward, increasing stakeholder awareness and engagement through electronic data collection methods will be crucial in capturing a more comprehensive picture and generating even more insightful results by using linear programming

## 7. AUTHOR CONTRIBUTIONS

Conceptualization, investigation, data curation, formal analysis, and writing of the original draft: Mohammed Kemal Ahmed. Reviewing, editing and supervision: Prof. Druga Prasad Sharma, Dr. Hussein Seid W: Advisor, Data curation and technical support: Prof. Achim Ibenthal, Technical support and fundraising: Dr. Getinet Yilma and Dharmveer Yadav.

## 8. FUNDING

Currently, Mohammed Kemal is a PhD scholar in software engineering at Addis Ababa Science & Technology University there is no separate fund

## 9. ACKNOWLEDGMENTS

I would like to acknowledge the Ministry of Agriculture for providing me with the data necessary to complete this study, and the agricultural office for granting me permission to visit their veterinary clinic frequently.

## 10. CONFLICTS OF INTEREST

The authors declare no conflict of interest.

## 11. DATA AVAILABILITY

With the permission of the Ministry of Agriculture of Ethiopia, the researcher collected the necessary data through an official letter from the University (Ref No: A/A/S/T./U/C/E /14/21). The researcher then visited different selected veterinary clinics in the country to accomplish the research

## 12. INFORMED CONSENT

There were no human subjects. / Since the research was conducted to propos the ELHRs in Ethiopia there were no human subject

## 13. ANIMAL SUBJECTS

While livestock research was conducted, as detailed in the data availability section, the data was gathered through an official letter request from the university to the Ministry of Agriculture, guaranteeing compliance with ethical guidelines and regulations. Additionally, informed consent was obtained from animal owners during data collection. Furthermore, animal handling was supervised by qualified veterinarian professionals.

## References

[1] Mwanga G, Mbega E, Yonah Z, Chagunda MGG. How Information Communication Technology Can Enhance Evidence-Based Decisions and Farm-To-Fork Animal Traceability for Livestock Farmers. ScientificWorldJournal. 2020;2020:1279569.

[2] Gizaw S, Desta H, Alemu B, Tegegne A, Wieland B. Importance of livestock diseases identified using participatory epidemiology in the highlands of Ethiopia. Trop. Anim. Health Prod. 2020;52:1745-1757.

[3] Erkyihun GA, Gari FR, Edao BM, Kassa GM. A review on One Health approach in Ethiopia. One Heal, Outlook. 2022;4:8.

[4] https://hdl.handle.net/10568/89973

[5] Gizaw S, Woldehanna M, Anteneh H, Ayledo G, Awol F, et al. Animal health service Delivery in crop-livestock and pastoral systems in Ethiopia. Front Vet Sci. 2021;8:601878.

[6] Cioffi R, Travaglioni M, Piscitelli G, Petrillo A, De Felice F. Artificial intelligence and machine learning applications in smart production: progress, trends, and directions. Sustainability. 2020;12: 492.

[7] Arega A, Sharma DP. Towards smart and green features of cloud computing in healthcare services: A systematic literature review. J Inf Syst Eng Bus Intell. 2023;9:161-180.

[8] Montagud S, Abrahão S, Insfran E. A systematic review of quality attributes and measures for software product lines. Softw Qual J. 2012;20:425-486.

[9] Hamid JS, Meaney C, Crowcroft NS, Granerod J, Beyene J. Cluster analysis for identifying sub-groups and selecting potential discriminatory variables in human encephalitis. BMC Infect Dis. 2010;10:364.

[10] El Idrissi AH. Digital technologies and implications for Veterinary Services. Rev Sci Tech. 2021;40:455-468.

[11] Ahmed MK, Sharma DP, Seid Worku H, Babu RB. Framework design for Machine Learning Integrated Mobile Based Livestock Disease Data Management, Diagnosis, and Treatment. J Surv Fish Sci. 2023;10:1482-1494.

[12] Beyene TJ, Asfaw F, Getachew Y, Tufa TB, Collins I, Beyi AF et al. A smartphone-based application improves the accuracy, completeness, and timeliness of cattle disease reporting and surveillance in Ethiopia. Front Vet Sci. 2018;5:2.

[13] Njenga MK, Kemunto N, Kahariri S, Holmstrom L, Oyas H, et al. High real-time reporting of domestic and wild animal diseases following rollout of mobile phone reporting system in Kenya. PLOS ONE. 2021;16:e0244119.

[14] Innocent GT, Vance C, Ewing DA, McKendrick IJ, Hailemariam S, et al. Patterns of mortality in domesticated ruminants in Ethiopia. Front Vet Sci. 2022;9:986739.

[15] Cichy C, Rass S. An overview of data quality frameworks. IEEE Access. 2019;7:24634-24648.

[16] Kim HY, Cho JS. Data governance framework for big data implementation with NPS Case Analysis in Korea. J Bus Retail Manag Res. 2018;12:36-46.

[17] Mositsa RJ, Van der Poll JA, Dongmo C. Towards a conceptual framework for data management in business intelligence. Information. 2023;14:547.

[18] Choi BCK. The Past, Present, and Future of PH Surveillance. Scientifica (Cairo). 2012:2012:875253.

[19] Falzon LC, Ogola JG, Odinga CO, Naboyshchikov L, Fèvre EM, et al. Electronic data collection to enhance disease surveillance at the slaughterhouse in a smallholder production system. Sci Rep. 2021;11:19447.

[20] George J, Häsler B, Komba E, Sindato C, Rweyemamu M, et al. Towards an integrated animal health surveillance system in Tanzania: making better use of existing and potential data sources for early warning surveillance. BMC Vet Res. 2021;17:109.

[21] Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinformatics. 2014;15; 1-11.

[22] Lin FR, Wu NJ, Tsay TK. Applications of cluster analysis and pattern recognition for Typhoon Hourly rainfall forecast. Adv Meteorol. 2017;2017:1-17.

[23] Saha R, Tariq MT, Hadi M, Xiao Y. Pattern recognition using clustering analysis to support transportation system management, operations, and modeling. J Adv Transp. 2019;2019:1-12.

[24] Hloušková Z, Lekešová M. Farm outcomes based on cluster analysis of compound farm evaluation. Agric Econ - Czech. 2020;66:435-443.

[25] Ishikawa S, Ikuta K, Obara Y, Oka A, Otani Y, et al. Cluster analysis to evaluate disease risk in periparturient dairy cattle. Anim Sci J. 2020;91:e13442

[26] Brock J, Lange M, Tratalos JA, More SJ, Graham DA, et al. Combining expert knowledge and machine-learning to classify herd types in livestock systems. Sci Rep. 2021;11:2989.

[27] Komaru Y, Yoshida T, Hamasaki Y, Nangaku M, Doi K. Hierarchical clustering analysis for predicting 1-year mortality after starting hemodialysis. Kidney Int Rep. 2020;5:1188-1195.

[28] Bustamam A, Fitria I, Umam K. Application of agglomerative clustering for analyzing phylogenetically on bacterium of saliva. AIP Conf Proc. 2017;1862.

[29] Almahamid F, Almahamid F, Grolinger K. Agglomerative hierarchical clustering with dynamic time agglomerative hierarchical clustering with dynamic time warping for household load curve clustering. In 2022 IEEE Can. Conf. Electr. Comput Eng. (CCECE). 2022:241-247.

[30] Ultsch A, Lötsch J. Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans). BMC Bioinformatics. 2022;23:233.

[31] Fedlu M, Seid A, Amana M. Review on community-based animal health workers in Ethiopia. Austin J Vet Sci Anim Husb. 2019;6:1-9.

[32] https://africacdc.org/download/framework-for-development-of-national-public-health-institutes-in-africa/

[33] Shiferaw J, Berhanu W. Livestock export from Adama quarantine stations: comparing management and biosecurity of feedlots before live animal export, DireDawa, Ethiopia. Adv Life Sci Technol. 2020;83:27-34.

[34] Praveen P, Kumar MR, Shaik MA, Ravikumar R, Kiran R. The comparative study on agglomerative hierarchical clustering using numerical data. IOP Conf Ser Mater Sci Eng. 2020;981.

[35] Januzaj Y, Beqiri E, Luma A. Determining the optimal number of clusters using silhouette score as a data mining technique. Int J Onl Eng. 2023;19:174-182.

[36] Gaido M. Distributed silhouette algorithm: evaluating clustering on big data. 2023. Arxiv Preprint: https://arxiv.org/pdf/2303.14102