International Journal of Computer Applications Technology and Research
Volume 15–Issue 11, 06 – 25, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1511.1002

Enhancing Infectious Disease Prediction in Slum Areas using a Sustainable Approach with Backpropagation Algorithm

Bernard Ochieng Okuku
Murang'a University of Technology
Department of Information Technology
Nairobi, Kenya

Geoffrey Mariga Wambugu
Murang'a University of Technology
Department of Information Technology
Nairobi, Kenya

Aaron Mogeni Oirere
Murang'a University of Technology
Department of Computer Science
Nairobi, Kenya

Abstract: Developing countries continue to grapple with the burden of infectious diseases in managing their public health. The growing use of technology and machine learning in healthcare has seen significant adoption of diagnostic and predictive ML models mostly in developed countries with enough supportive resources. For resource-constrained countries with substantial populations living in informal settlements, there has been a growing need for the integration of ML in their healthcare management. In response, we design a sustainable machine learning model for improving the prediction of infectious diseases in informal settlements. We strive to address limitations of existing predictive models, particularly their inability to effectively handle noisy, incomplete, and resource-constrained datasets common in slum environments. The research is handled in phases with this paper covering the first objective which aims at understanding the prevalence of infectious diseases and existing management strategies in informal settlements of Kibera, Mathare, and Mukuru kwa Njenga in Nairobi, Kenya. To achieve this, we adopt a mixed-methods research design to investigate the distribution and management of infectious diseases in selected areas. The subsequent phases of the research will evaluate existing machine learning datasets, and finally develop and evaluate a setting-specific ANN-based predictive model. The model will be trained using the backpropagation algorithm with stochastic gradient descent optimization and evaluated using standard machine learning performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis. In the first phase, the paper confirms high prevalence of Tuberculosis, Malaria, Cholera and HIV/AIDs, as well as existing management strategies adopted and the challenges in informal settlement. The proposed framework is expected to improve predictive accuracy by learning complex nonlinear relationships between clinical symptoms, environmental conditions, and socio-demographic factors. While we acknowledge that infectious diseases remain a serious health challenge in informal settlements, we believe that the proposed model provides a promising approach for developing sustainable, scalable, and cost-effective infectious disease prediction systems tailored to informal settlements. Future implementation should further validate the model using real-world clinical and community datasets, integrate it with existing health information systems, and continuously retrain it using updated epidemiological data. Policymakers should also support digital health infrastructure and standardized data collection to enhance predictive healthcare in informal settlements. This research contributes to the growing application of artificial intelligence in public health by demonstrating how optimized backpropagation-based neural networks can support sustainable infectious disease prediction in resource-constrained environments.

Keywords: Infectious disease prediction, backpropagation neural network, informal settlements, sustainable disease surveillance, machine learning in public health

1. INTRODUCTION

Infectious diseases are viral or bacterial infections transmitted from one person to another through airborne viruses or contact with bodily fluids and blood [1]. These diseases are prevalent in slum areas and account for a significant number of fatalities especially due to poor management. While significant strategies have been deployed to help curb their spread, critical gaps still exist making these areas high risk. Novel solutions are thus needed to complement existing techniques and facilitate efficient management of these infections. Currently the identification of some of these diseases like TB is largely carried out through direct visual inspection. However, in the current era, they can be diagnosed by means of digital image processing. This paper recommends an expert system model using neural network approach to facilitate the prediction of the specific diseases. Similar works [2, 3] recommend the use of backpropagation and gradient descent optimization techniques together with artificial neural networks when designing such models. This enhances the prediction of the disease by the system according to the symptoms entered.

Several studies highlight a myriad of challenges in managing infectious diseases in slum areas. Research carried out by other scholars [4,5] point out overcrowding and poor sanitation, high population density, lack of resources, and restricted access to healthcare as some of the challenges that facilitate the rapid spread of infectious diseases, making it difficult to control outbreaks and implement effective management strategies. Either, another study [6] found that poor infrastructure in these areas, including unreliable electricity and water supply, the lack of proper health record and reporting systems, impede the functioning of diagnostic equipment, the storage of medical supplies, as well as accurate reporting and tracking. Likewise, similar studies [1]

have confirmed socioeconomic barriers where poverty, illiteracy, and lack of awareness about health issues have been found to prevent residents from seeking medical help, leading to underreporting and late detection of diseases. In his work, Tadesse [4] also found that environmental factors characterized by poor sanitation, stagnant water and waste accumulation, provide breeding grounds for vectors like mosquitoes, further complicating disease prediction and control efforts. Others like Bhalla and Aggarwal [7] further indicate that cultural beliefs and stigma associated with certain diseases can discourage individuals from reporting symptoms or seeking treatment, leading to delayed diagnosis and intervention. Residents of these areas have also shown notable resistance to interventions due to mistrust of authorities or misinformation, further limiting efforts to predict, diagnose, and control infectious diseases effectively [8].

To help address these challenges, this paper develops an Artificial Neural Network (ANN) with a Backpropagation algorithm to enhance prediction. ANN is a computerized model whose structure and functionalities are similar to the biological neural networks that exist in the human brain, which are known to process information [9]. According to [10], ANNs are used in artificial intelligence and specifically in machine learning to recognize patterns, make decisions, and predict outcomes based on input data. This approach uses gradient descent for supervised learning of ANNs as applied in similar works [11].

In the context of this study, we consider sustainability as the development of a disease prediction system that can operate effectively and continuously in slum areas using minimal resources, while being technically reliable, economically affordable, socially acceptable, and environmentally considerate. This ensures that the model and its supporting infrastructure can be maintained, scaled, and adapted over time without heavy dependence on external funding or advanced technology. We therefore emphasize the long-term use of low-cost technologies, local capacity, minimal resource consumption, and social acceptance, ensuring that the system can function effectively without creating dependency or requiring significant ongoing external support. The algorithm can help in achieving cost-effective implementation leading to efficient resource allocation. Automated diagnosis using neural networks can reduce the reliance on expensive laboratory tests and medical experts, providing affordable healthcare solutions to impoverished communities [9, 11]. By predicting disease outbreaks and identifying high-risk areas, resources such as vaccines, medications, and healthcare personnel can be allocated more efficiently, minimizing wastage and maximizing impact.

We further believe that such a sustainable model will help in solving challenges that arise in integration with existing health systems. This is supported by evidence from similar works [7, 9, 11] that have shown that neural networks can integrate with existing health information systems, improving data sharing and coordination among healthcare providers, NGOs, and government agencies. The proposed algorithm will also provide support for public health strategies. Insights generated from these models can support public health strategies, such as vaccination campaigns, hygiene education, and environmental improvements, tailored to the specific needs of slum areas.

1.2. Problem Statement

Traditional machine learning models mostly use Support Vector Machines (SVM) and Naïve Bayes algorithms for classification and diagnosis of diseases [12]. Others employ certainty factors and forward chaining mechanisms, which have lower diagnostic accuracy and are more prone to errors [14]. In [13], an attention model was incorporated into the Long Short-Term Memory (LSTM), as it has been demonstrated that adding such a model significantly enhances the LSTM's performance in other application domains. Though powerful, these models often struggle with the complex and noisy data prevalent in slum areas, which limits their predictive accuracy and sustainability.

We also observe that the above existing predictive models for infectious diseases have been primarily applied in urban areas of developed countries with better living conditions. There is no predictive model specifically designed for informal and resource-constrained areas in developing countries like Kenya. Current models also are about 94% accurate [14], leaving room for erroneous diagnosis. They also struggle with the noisy, complex data in slums, which limits their predictive accuracy and sustainability. Moreover, there are high levels of ignorance, skepticism and resistance among people, including patients and doctors, regarding the accuracy and reliability of automated systems for medical diagnosis, which affects the adoption of such systems in medical practice.

To address these issues, there is a need for an improved system that can accurately predict infectious diseases according to their symptoms within informal environments. Our proposed predictive model forecasts disease outbreaks based on various setting-specific indicators, helping in proactive measures and resource allocation. Enhancing the ANN model with backpropagation algorithm, and optimizing through gradient descent technique, can significantly reduce errors currently observed. The optimization will enable the algorithm to handle the unique data challenges of slums, such as high variability, incomplete datasets, and real-time processing needs [10]. By comparing the accuracy of training and test sets for various diseases, ANN models can achieve higher accuracy with minimal errors and cost functions. This optimization will not only improve medical diagnostic systems but also promote the integration of technology in healthcare, contributing to better disease management in low-resource settings.

1.3. Research Objectives

1.3.1. Main Research Objective

The general objective is to develop a machine learning model for a sustainable prediction of infectious diseases in informal settlements by utilizing the backpropagation algorithm.

1.3.2. Specific Research Objectives

i. To establish the distribution and the existing

management strategies of infectious diseases in the informal settlements.

ii. To analyze the existing machine learning datasets so as to determine their suitability in modelling prediction of infectious diseases in the informal settlements.

iii. To design a machine learning model for prediction of infectious diseases.

iv. To test and evaluate the machine learning model performance based on the selected features in the prediction of infectious diseases in the informal settlements.

1.4. Scope of the Study

The scope of this paper is limited to Objective (i) and in particular, to the prediction of the infectious diseases in the informal settlements of Kibera, Mathare, and Mukuru kwa Njenga areas of Nairobi, Kenya. The infectious diseases considered in this study include Tuberculosis (TB), Malaria, HIV/AIDS and diarrheal diseases such as cholera. The complete approach will involve developing an Artificial Neural Network prediction model that is scalable, cost-effective, and resource-efficient, tailored for low-resource settings in slum areas using backpropagation algorithm.

1.5. Limitations of the Study

This research faces some limitations related to data, methodology, and context. Data-related challenges include the unavailability or poor quality of health, environmental, and geospatial data in slum areas, where healthcare systems are often informal and fragmented. Missing or noisy data, inconsistent reporting, and a lack of longitudinal datasets can also hinder the accuracy of the predictive model. Furthermore, the findings may not be as broadly applicable to other areas due to the diversity of slum communities and their distinct socioeconomic and environmental characteristics.

From a technical perspective, the computational requirements of the backpropagation algorithm may exceed the resources available in slum settings, potentially affecting the scalability of the solution. The study also faces challenges in balancing the need for a simple, sustainable approach with achieving high predictive accuracy. The interpretability of neural networks and other machine learning models, could limit their practical adoption by healthcare workers and policymakers. Moreover, rapid demographic and environmental changes in slum areas, as well as external factors like natural disasters or political instability, may affect the applicability and validation of the model over time. Despite these limitations, the study acknowledges them as opportunities for future research and emphasizes the need for collaboration with stakeholders to overcome these challenges effectively.

2. RELATED WORKS AND BACKGROUND

This section primarily reviews related works on the relevant areas of this research. It provides scholarly understanding of infectious diseases and management strategies within informal settlements. Drawing from existing literature, we also discuss the emerging role of machine learning ANNs data-driven diagnostic and predictive approaches used in handling infectious diseases.

2.1. Distribution of Infectious Diseases in Informal Settlements

A systemic review [6] has shown that residents of informal settlements are at disproportionately higher risks for infectious diseases compared to non-slum urban residents, primarily due to persistent overcrowding, poor ventilation, and limited access to quality healthcare. Informal settlements have also been documented to have elevated mortality and morbidity from respiratory and diarrheal infections relative to wider urban populations, a clear indication of increased disease clustering in dense informal settings [1]. In Kibera, for instance, metagenomic analyses of sewage [2] identified the presence of multiple bacterial, parasitic, and viral pathogens over time, illustrating elevated risks of contamination in informal settlements. These findings align with the claims that distribution of infectious diseases in urban informal settlements is both spatially concentrated and temporally varied.

Residents of informal settlements have been found to have complex demographic patterns of risks of contamination. Surveillance data from Nairobi slums [7], for instance, shows that tuberculosis constitutes a significant proportion of adult mortality burden, further pointing to the broader systemic vulnerabilities associated with informal settlements. Likewise, a study [12] carried out in Mathare on diarrheal infections has associated environmental and sanitation practices with disproportionate risks of contamination for households with limited access to proper sanitation, especially for children under the age of five. These patterns indicate the effect of demographic factors like living spaces, age, water, sanitation, and hygiene conditions on the distribution of infectious diseases in informal settings.

2.2. Management Strategies for Infectious Diseases in Informal Settlements

2.2.1. Formal Health System Utilization

Other works have equally documented the use of a mix of formal and informal healthcare options in the utilization of healthcare services. For instance, a classic healthcare-use research conducted in Kibera found that whereas a significant population reported cases of illnesses at some point, majority did not seek formal healthcare services instead opting for over-the-counter or informal care, citing challenges with accessibility and affordability [8]. This is an indication of limited access and utilization of quality formal healthcare system in these settings despite the high infectious disease burden.

2.2.2. Community and Household-Level Management Practices

Community-level disease management strategies have been

adopted to help address challenges with accessing formal healthcare. In their work [13], the scholars propose that community-based interventions focusing on CHW engagement, health education, and community screening, among others, can be effective in the prevention and management of infectious diseases such as Hepatitis A within the urban low-income settings. These techniques, however, vary in reach and implementation. Jahnavi and others [14] further illustrates the linkages between sanitation barriers, gender dynamics, and daily lifestyles adopted in the management of COVID-19 and behavioural practices that influence health outcomes in settings with poor sanitation infrastructure.

2.2.3. Barriers to Treatment Completion and Effective Management

The barriers to effective management of infectious diseases in informal settlements are multidimensional. Evidence has shown that stigma and limited continuity of care, in addition to economic and geographic barriers discussed above, affect adherence and completion of treatment. A study [5] in Kenya's informal settings suggest that improved sanitation and hygiene efforts are correlated with both nutritional status and risks of infection among children below 5 years, further suggesting that integrated health and environmental strategies are crucial in enhancing the management outcomes of infectious diseases in these areas.

2.3. Emerging Role of Data-Driven and Predictive Approaches

Empirical research [2], proposes the potential of integrating new surveillance techniques like metagenomic pathogen detection with conventional epidemiological approaches to improve detection and response to infectious agents in informal settlements. While in the recent times, predictive modelling and machine learning have been increasingly proposed to help in public-health management, critics [11,12] caution that models must be grounded in empirical data that reflect lived conditions in targeted settings to minimize misclassification and equity aspects.

2.4. Artificial Neural Networks and Backpropagation Algorithm

Among the deep learning models, the ANN model has certain characteristics that are almost identical to those of the human brain [9]. The neurons that make up the human brain are linked to one another and are responsible for processing information. ANN also contains neurons, which are artificially generated computer processing elements. These neurons are basically what constitute the synthetic structure of an ANN.

The three primary layers of an artificial neural network comprises of the input layer, the hidden layer, and the output layer [9]. The input layer, which is the initial layer, accepts certain parameters as input. The output layer, which comes last, displays the results of the processing completed in the preceding layer, whereas the hidden layers are the intermediary layers. Multiple hidden layers may exist in an ANN, and experts find that the more hidden layers a model has, the deeper it becomes and the greater its capability for learning [4]. These layers are connected to each other through "weights" which governs the learning of the model. Training an ANN is achieved by modifying and fine-tuning the weights in order to address a specific problem [2]. Each neuron is activated by applying one of the several activation functions as shown in figure1.

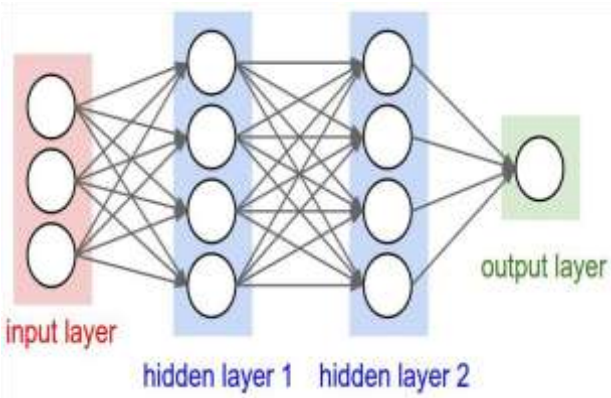

Figure.1. ANN Model [**Error! Reference source not found.**]

During the forward propagation process, which occurs during ANN training, all inputs are sent into the neural network. The expected and the actual outputs are then compared. Once the feed-forward phase is complete, the ANN itself computes the errors [2]. The network can be trained by adjusting the weights because these errors are backpropagated in the opposite direction [17]. This suggests that backpropagation must be carried out enabling users to update all of the weights at once.

2.4.1. Challenges of Backpropagation Algorithm

When using the backpropagation algorithm, Gibson et al. [16] admit that several open challenges still exist that need to be addressed. Some of these challenges include vanishing and exploding gradients [18], slow convergence [19], overfitting [20], computational cost and sensitivity to initialization [21]. Other experts [12, 16, 22, 23, 24, 25] have also pointed out more challenges, including local minima and saddle points, data scaling and preprocessing requirements, hyperparameter tuning, lack of interpretability, scalability issues and high memory usage.

2.5. Gradient Descent Optimization Technique

The backpropagation algorithm backpropagates the error and updates the weights through the neural network [12]. Backpropagation is a brute-force method for minimizing the error and cost function, therefore while it minimizes error, it does not optimize. This situation may give rise to the curse of dimensionality. Backpropagation and an improved optimization approach are therefore required in order to improve the network performance and eliminate errors in a more accurate and timely manner [7]. This can be solved by

use of the Gradient Descent Optimization technique whereby the cost function is minimized by repeatedly progressing in the direction of the steepest descent, which is indicated by a negative gradient [19]. Scholars like Tadesse [4] generally affirm that gradient descent is a reasonable optimization method; nevertheless, it may find the local minimum of the cost function and overlook the global minimum when the curve is asymmetrical and has several minimum points, which is unsatisfactory. In these circumstances, the system cannot be fully optimized. We thus suggest the use of the Stochastic Gradient Descent Optimization technique to solve this problem.

2.5.1. Stochastic Gradient Descent

In the Stochastic Gradient Descent (SGD) technique, the parameters are updated one by one for every training instance. Therefore, depending on the problem, the batch gradient dropdown is slower than the SGD. Regular updates provide a reasonably detailed rate of development, which is advantageous [7]. The problem is that, when considering a slow descent to batch, frequent updates are more costly from a computational perspective. Instead of gradually lowering the error rate, Hossain et al [26] warn that the frequency of these updates may instead produce noisy gradients that raise it. We shall therefore integrate ADAM, which is an SDG approach, with our backpropagation method.

2.6. Machine Learning Model Development

A machine learning (ML) model is a program that can analyze a previously unknown dataset to identify patterns or draw conclusions [26]. These kinds of tasks can be performed by a machine learning model that has been "trained" using a very large dataset. Depending on the task, the model is trained to recognize specific outputs or patterns from the dataset. This approach subsequently generates a model with specific rules and data structures.

Machine learning has become increasingly popular in infectious disease prediction. It has been used in extracting valuable insights from medical data thus promoting prediction [12]. Developing an ML model involves collecting data from several reliable sources, processing the data to make it suitable for model development, choosing an algorithm, building the model, calculating performance metrics, and choosing the model with the best performance [4].

In the ML model development lifecycle, these procedures are grouped under the broad categories of data exploration, model development, model hyperparameters modification, and the best-performing model selection [7, 27]. Exploratory data analysis is an essential first step after the business hypothesis has been prepared. The phase requires up to 40-50% of the total study time since the model's conclusion relies on the quality of the input data used to train it. According to the documentation by TensorFllow [20], finding data characteristics, preparing data, and feature engineering are all part of exploratory data analysis. Determining the target/class variable (output), predictor/features variables (inputs), and data types (string, numeric, or datetime) are all steps in the attributes identification process. Additionally, the features must be categorized into continuous and categorical variables [27]. This facilitates the application of appropriate variable treatment by the algorithm during the development of the model. To fill in the gaps and improve the prediction power of the model, identification of the missing values and outliers is done during data pre-processing, and the mean or median for quantitative features and the mode for qualitative qualities are computed. As further confirmed by Bhalla and Aggarwal in their previous work [7], it is possible to eliminate the outliers that raise the mean and standard deviation by calculating the natural log value, which reduces the volatility caused by extreme values.

When developing an ML model, the data must be split into two separate sets. These include a "training set" and a "testing set," in the ratio of 80:20 or 70:30 [28]. A range of supervised algorithms [designed for labelled data] and unsupervised algorithms [designed for unlabelled data] is available for selection. This is determined by the nature of the input data and the specific medical outcome being predicted.

Once the model is developed, it is important to maintain it to prevent it from becoming outdated over time. Bhalla and Aggarwal [7] recommend that it is crucial to keep it updated and applicable in response to modifications in the source data. Additionally, as the number of models increases, proper management of their configuration becomes crucial to their overall performance [2]. The entire process that an ML model undergoes from the identification of source data to model construction, deployment, and maintenance is referred to as the model lifecycle. Figure 2 illustrates the phases involved in the ML model development.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 11, 06 – 25, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1511.1002

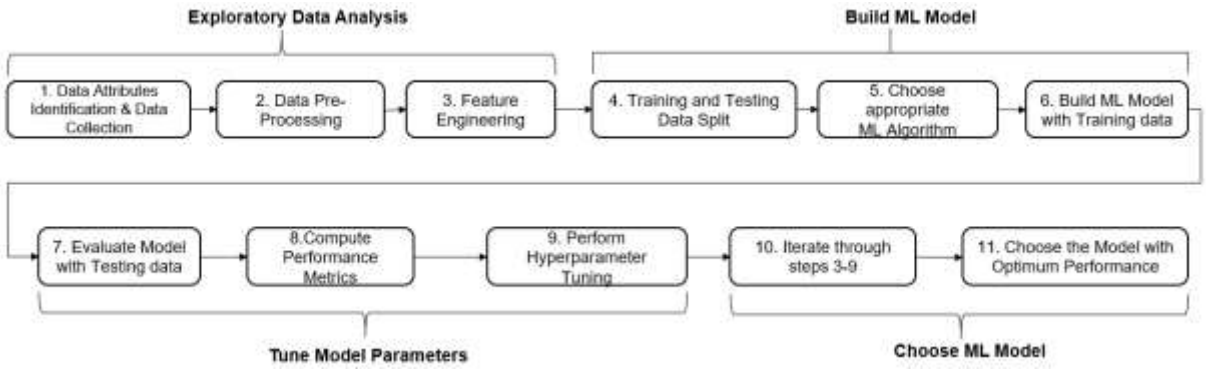

Figure.2. Machine Learning (ML) Model Development Lifecyle [**Error! Reference source not found.**]

2.7. Limitations of Existing Prediction Models

In their study, Fathima and Hundewale [12], employs Support Vector Machines (SVM) and Naïve Bayes approaches to develop their system. This approach is designed exclusively for the diagnosis and categorization of dengue, and its accuracy percentage is relatively lower than that of other systems. SVMs are an effective class of supervised learning techniques for regression, classification, and outlier identification, and their utility with various kernel functions and efficacy in high-dimensional spaces are especially well-known.

Another work [13] utilizes certainty factors and the forward chaining processes. The researcher, however, points out that these techniques have a lower diagnostic accuracy and are more prone to errors. Although it has been demonstrated that adding an attention model to the LSTM significantly enhances its performance in other application areas, the impact of doing so was, however, not examined by Subiyanto et al. [15].

Unfortunately, in the study by Hossain et al. [26], the validation cohort lacked certain attributes from the discovery cohort dataset, and vice versa. Additional parallels or differences between the resulting feature rankings may have been discussed if the clinical variables in both datasets had been identical. The researchers were unable to use the same methodology for hepatitis C and cirrhosis bioinformatics datasets, which included microarray gene expression data.

Additionally, the study did not employ strategies to address data imbalance or take into account alternate data imputation approaches as other works do [29, 30].

2.8. The Health Belief Model (HBM)

2.8.1. Overview of the Health Belief Model

The Health Belief Model (HBM) is among the known and widely adopted theoretical frameworks in health-related studies particularly in developing deeper understanding on health-related decision-making and behaviors. Conventionally designed to explain preventive health actions, the model argues that individual's involvement in health behaviors is shaped by their subjective assessment of associated risks, perceived effects, and perceived value of the remedies taken to mitigate those risks. This aspect has made the model increasingly adapted to infectious disease contexts over time, with tuberculosis management, where behavioral responses are critical in prevention, detection, and treatments adherence being among the key examples. Figure 3 presents the traditional formulation of the HBM, where the model conceptualizes health behavior as a function of six core constructs which include perceived susceptibility, perceived severity, perceived benefits, perceived barriers, cues to action, and self-efficacy, which together operate within a broader demographic and psychological context [4,15].

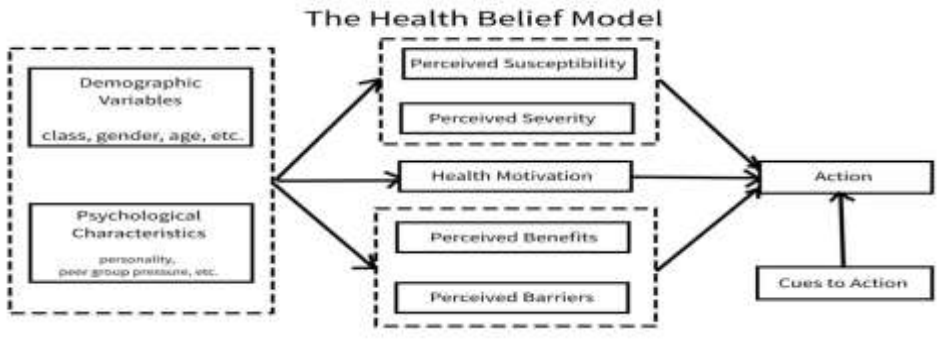

Figure 3. Traditional Health Belief Model

2.8.2. Understanding Traditional Health Belief Model Constructs

Perceived susceptibility denotes one's subjective assessment of their likelihood of contracting an infection. Within the context of infectious diseases, this may include perceived exposure risks through getting contact with infected individuals, social settings or environmental conditions. According to documented evidence from recent studies [26], perceived susceptibility has significant influences on health-seeking behaviors like regularly screening for TB, and early diagnosis particularly in overcrowded urban settings. In informal settlements such as Kibera, Mathare, and Baba Dogo, susceptibility perceptions are influenced by overcrowding, communal sanitation facilities, and shared living spaces, which intensify awareness of risks of exposure, biomedical knowledge limitations notwithstanding.

Perceived severity refers to the beliefs around the seriousness of an infection and related potential consequences, such as physical, social, and economic effects. In the case of infectious diseases, for example, severity perceptions often involve fear of prolonged illness, social stigma, income loss, and even mortality. Recent studies [18] confirm that the higher the perceived severity, the higher the likelihood of seeking early diagnosis and adhering to treatment regimens. In low-income households, however, it has been identified that severity perceptions may coexist with normalization of illness or fatalism, which may paradoxically delay seeking care [1].

Perceived benefits reflect beliefs regarding the effectiveness of recommended strategies in mitigating infection risks or severity. For infectious diseases such as tuberculosis, this may include confidence in diagnostic services, trusting treatment efficacy, and belief in the benefits of early diagnosis. Empirical evidence [10] suggests a strong predictive relation between perceived benefits and tuberculosis screening uptake and treatment adherence, especially in settings where healthcare is perceived as accessible and reliable.

Perceived barriers highlight both real and perceived barriers that constrain engagement in health behaviors. These may encompass factors such as financial costs, proximity to healthcare facilities, stigma, opportunity costs, and fear of discrimination. In informal settings, barriers are often structural and dynamic, further exacerbating risks of delayed diagnosis and incomplete treatment [8, 10]. Evidence has shown that perceived barriers is among the strongest predictors of non-adherence to tuberculosis treatment, even when perceptions on awareness and severity are high.

Cues to action are basically triggers that activate individual's readiness to engage in health behaviour. These may be both internal for example symptoms, and external such as CHW visits and health campaigns among others. Previous works [4, 25] propose targeted community-based cues like door-to-door screening and peer referrals, as crucial approaches to improving tuberculosis case detection in urban informal settlements.

Self-efficacy is individual's confidence in their ability to successfully perform a health-related action. In tuberculosis management for instance, this influences the willingness to undergo prolonged treatment and faithfully adhere to medication regimens. Research targeting urban slums like that by Gulis et al. [1] suggest that social support and prior successful treatment experiences significantly improve self-efficacy and treatment completion rates.

As illustrated in Figure 3, demographic variables such as age, gender and socioeconomic status, and psychological characteristics like social norms and peer influence modify the core HBM constructs. These modifiers, however, do not directly determine behaviour, rather they shape individual's perception of risks, benefits, and barriers. Showkatian and others [10] emphasizes the need to contextualize HBM constructs within the structural determinants of health, especially in urban slam areas.

2.8.3. Limitations of the Traditional HBM in Infectious Disease Management

Whereas the traditional HBM has demonstrated significant explanatory power, the model has notable limitations in complex infectious disease environments. In particular, the model treats construct largely as linear and independent, relies on self-reported perceptions, and lacks mechanisms for capturing non-linear interactions among multiple risk factors. Such limitations are specifically salient in infectious disease transmission dynamics within informal settlements, where environmental, behavioral, and structural factors interact in complex, non-linear manner.

2.8.4. Adaptation of the Health Belief Model for Infectious Disease Prediction

To help address the limitations of traditional HBM as discussed, this study adopts an adapted Health Belief Model integrated with a backpropagation neural network as shown in Figure 4 below.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 11, 06 – 25, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1511.1002

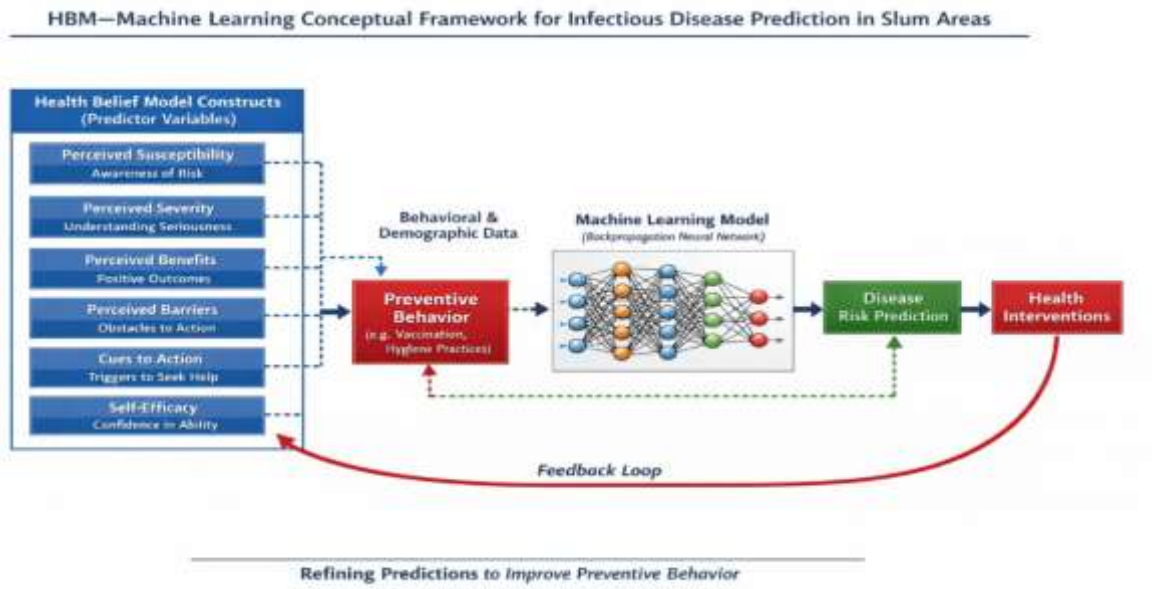

Figure 4. Adapted Health Belief Model for infectious disease prediction [Rosentock]

2.8.5. Conceptual Rationale for Adaptation

The adaptable model reconceptualizes the traditional HBM constructs as quantifiable features rather than purely explanatory variables. It operationalizes perceived susceptibility, severity, benefits, barriers, cues to action, and self-efficacy, suing measurable indicators captured through the survey instruments incorporating factors like crowding, diagnosis access, and treatment history among others as discussed in later section. Existing novel methodological works encourage this transition from theory-driven behavioral construct to data-driven predictive representations, particularly in infections disease prediction modelling [25].

2.9. HBM Integration with Backpropagation Neural Networks

In the adapted model as presented in Fig. 4, HBM constructs form the input layer of a backpropagation neural network. The hidden layers model creates nonlinear interactions between behavioural perceptions, environmental exposure, and health system access. The output layer generates probabilistic tuberculosis risk scores at individual or household levels. This way, it is agreed that the model is able to capture complex interactions overlooked by linear models, learn from empirical data rather than predefined assumptions, and generate actionable outputs like hotspot identification and targeted screening priorities [9].

2.10. Relevance of the Adapted HBM to this Study

Grounding this study in the Health Belief Model ensures that predictive modelling remains anchored in established behavioral theory. The adapted HBM provides a conceptual linkage between public health theory and machine learning, ensuring that infectious disease risk prediction is both computationally accurate, behaviorally interpretable and policy relevant.

2.11. The Conceptual Framework

The primary independent, moderating, and dependent variables are described here. Independent variables include Sanitation Facilities (SF), the Patient Gender (PG), number of infectious diseases management strategies employed (NMS), Detection Time (DT), Number of Detection Practices (NDP) and Access to Healthcare (AH). The Moderating variables are Health workers' gender (G) and age (A), health workers' education level (E), residential ownership (L), the chosen model (M), infectious diseases pictures collected (D), quality sanitation variables collected (Q). Dependent variables (Yx) include distribution and intensity of infectious diseases (Yi), accuracy of infectious diseases diagnosis (Yii) and accuracy of infectious diseases prediction (Yiii).

International Journal of Computer Applications Technology and Research
Volume 15–Issue 11, 06 – 25, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1511.1002

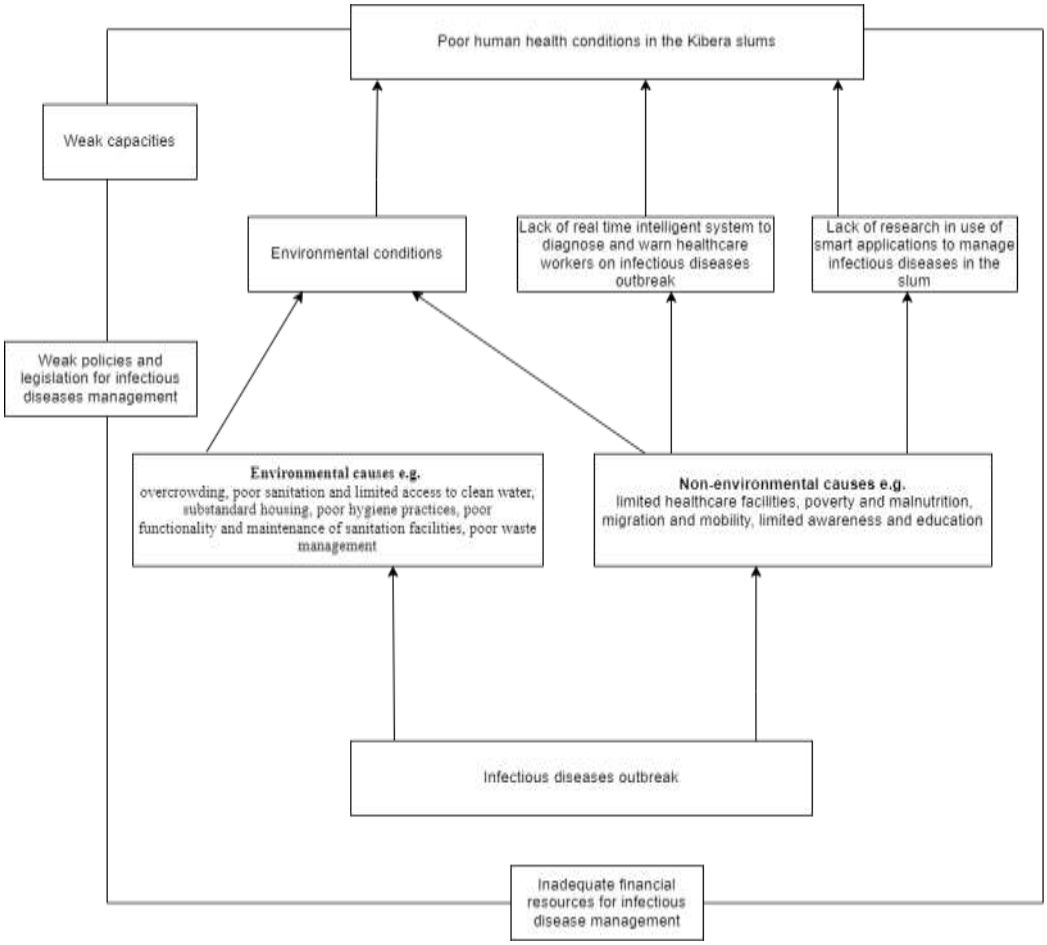

Figure 5. Conceptual framework

3. METHODOLOGY

We discuss the methodology used in the study in this section, and outline the approaches used to achieve the intended objectives. We present how the research was designed and discuss the techniques used to achieve the results.

3.1. Research Design and Approach

The objective of the research design is to give a proper framework for the study. In response to the first objective, this paper adopts a mixed methods research design. To establish the distribution and the existing management strategies of infectious diseases in the informal settlements, the study undertakes a review of existing literature and carries out a survey with the population in the informal settlements. Through this approach, we aimed to determine the nature, extent of distribution and existing knowledge of infectious diseases in the informal settlements in relation to the sanitary conditions. It also establishes tools used in the management of infectious diseases. This includes factors that determine the distribution of cases of infectious diseases. To monitor this objective, field reports, Electronic Clinical Records and survey findings are further analyzed.

In the second objective, the study will thoroughly analyze the existing and accessible machine learning datasets from secondary sources so as to determine their suitability in modelling prediction of infectious diseases in the informal settlements. The outcome of this objective will further guide the selection of the training dataset for the proposed model for objective three. To meet the third objective, the study will adopt the design science research methodology suggested by Peffers et al. [31] to design an effective machine learning model for prediction of infectious diseases. The fourth objective involves testing and performance evaluation of the model designed in objective 3 using the features chosen and dimensionality reduction in predicting infectious diseases in the informal settlements.

3.2. Data Collection

Data was sourced from both primary and secondary sources. The primary data were obtained through survey research. Structured surveys and questionnaires administered through Kobo forms were used to collect standardized data. Participants were chosen by Purposive Sampling based on infectious diseases and informal settlements characteristics. This ensured that the sample included individuals who have the necessary knowledge and experience to provide valuable insights on the subject. Snowball sampling was equally applied. After identifying the key health informants, they were able to recommend other health workers who met the study criteria. Multiple-choice, Likert scale, and a few open-ended

questions for rapid qualitative insights were all included in the questionnaires. Epidemiological data were also collected with the aim of gathering data on disease prevalence and the types of diagnoses commonly carried out in these areas. This involved collecting data from clinical records, patient registers, and diagnostic reports.

For subsequent objectives, additional large data to validate the model will be required and thus, collection of secondary data of people with infectious diseases' images dataset from existing sources for computer image processing will involve use of programming (Python) libraries to collect images from medical web-based sites. Other datasets will be obtained from publicly accessible reputable sources.

3.3. Research Population and Sample Size

3.3.1. Research Population

Residents of informal settlements of Kibera, Mathare, and Mukuru kwa Njenga within Nairobi City County were the main target population for this study. This population is defined by geographic, socioeconomic, and health-related characteristics. This study also involved interacting with a number of stakeholders including healthcare providers such as doctors, nurses, Community Health Workers (CHWs), pharmacists, and laboratory technicians among others.

3.3.2. Sample Size

For quantitative data, a total of 46 household were covered across the three selected settlements to ensure statistical power and representativeness. We further conducted 15 in-depth interviews with key informants such as healthcare workers to gather qualitative insights. Additionally, 3 focus group discussions, each with 6–10 participants were carried out to capture community perspectives and themes.

3.4. Data Pre-processing

3.4.1. Data Cleaning

Identification and imputation were used to handle missing data. Missing values which might be represented as NULL, blanks, or placeholders like "NA" were identified. Depending on the extent and nature of missing data, the study chose an appropriate imputation method. We used the mode for categorical data and the column mean or median for numerical data to impute missing values. The removal of the impacted rows or columns was considered if the missing data was widespread and could not be properly imputed, but only if doing so did not result in a major loss of information. The study looked for and eliminated duplicate records, which may have arisen from repetitive data gathering or mistakes in data entry. Any errors encountered were corrected either through validation or manual correction. The validation of the data entries was done by cross-referencing with reliable sources or using logical checks, for instance, ensuring ages were within a reasonable range. Obvious errors, such as typographical mistakes were also corrected.

3.4.2. Data Transformation

Standardized numerical features either had a mean of 0 or a standard deviation of 1, usually between 0 and 1. This was especially crucial when different features had varying scales, like age in years. The study also applied logarithmic or other transformations to skewed data to normalize distributions. Categorical variables and labels were encoded using One-Hot Encoding and Label Encoding. Label Encoding was used to encode categorical labels by transforming them into numerical values, such as "Male" = 0 and "Female" = 1. One-Hot Encoding was used to generate binary columns for every category in order to prevent ordinal associations for categorical variables that included more than one category, such as dwelling houses.

3.4.3. Model Training Dataset and Dataset Preparation

While the 46-household dataset provided epidemiological distribution and management context, an independently documented secondary large machine-learning dataset will be adopted upon thorough analysis of its quality and quantity in objective (ii). The essence of this separation is to prevent the use of this relatively small exploratory survey dataset in training a high-dimensional predictive model as well as providing a substantially larger sample for model development. This household dataset will further be essential in guiding the selection of the training dataset as it provides information critical for variable selection ensuring the model developed remains both useful and relevant to the local informal settings. Key variables to be considered will therefore include demographic, clinical history and TB-related clinical information as well as any other characteristics relevant to TB prediction. The dataset will be divided in the 70:30 ratios for training and testing data respectively meeting the requirement to preserve an independent hold-out sample for final evaluation.

3.4.4. Feature Selection

Feature selection will encompass both empirical findings from the study outcome, clinical relevance and statistical screening. As presented in the results section of this study, the descriptive and inferential analyses identified variables with significant association with disease occurrence. This will be critical in screening variables in the external dataset selected for model training for completeness, redundancy, and their predictive relevance.

3.5. Model Architecture

The proposed predictive model will be implemented as a fully connected feed-forward ANN trained with backpropagation. The input layer will comprise six predictor nodes corresponding to the selected variables as per the selection criteria above. Two hidden layers will be specified to capture nonlinear relationships. The first hidden layer will have 12 neurons, and the second layer 6 neurons. Sigmoid activation will be used in these hidden layers whereas a single sigmoid output neuron will generate the probability of the disease

International Journal of Computer Applications Technology and Research
Volume 15–Issue 11, 06 – 25, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1511.1002

occurrence. Since the dependent variable is binary, binary cross-entropy will be used. Figure. 6 is a data flow diagram that depicts the processing and data flow through the system.

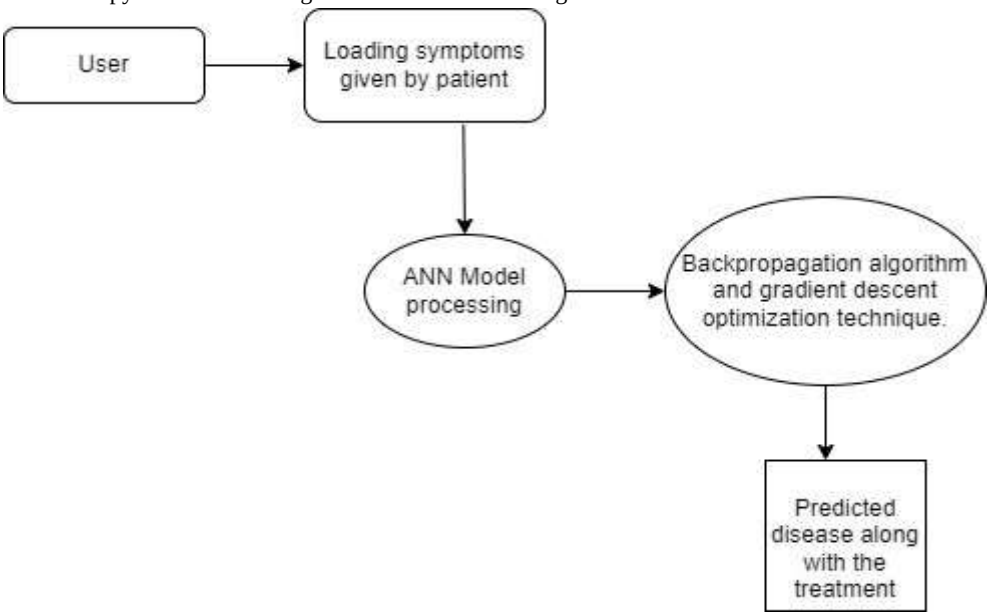

Figure 6. Data Flow Diagram of the Model

3.5.1. The Adopted HBM-SBPNN Model Architecture

An overview of the model architecture summarizing the links between the study objectives is presented in Fig. 7. It outlines the seven interconnected stages, starting with the relevance of the results in this paper, the existing ML dataset evaluation that follows, the actual model development, and ending with a validated sustainable infectious disease prediction ANN model as expected.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 11, 06 – 25, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1511.1002

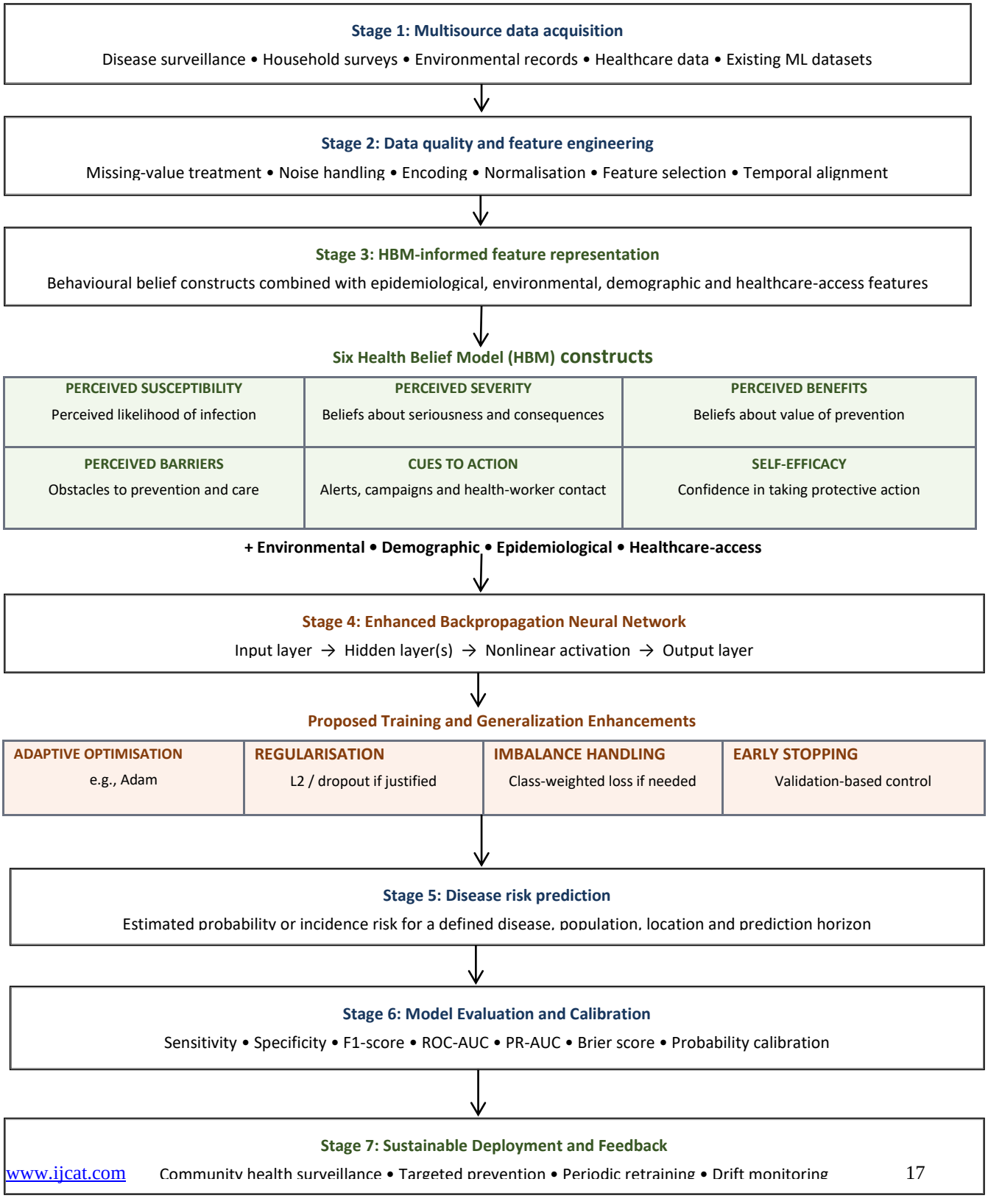

Figure 7. Adopted HBM-SBPNN model architecture

3.5.2. Hyperparameter Tuning and Training

Hyperparameters will be optimized using grid search within the training data and evaluated through k-fold cross-validation. This will comprise search space of 0.001 and 0.01 rates, 8 and 16 batch sizes, and Adam and stochastic gradient descent (SGD) optimizers. Given the potential for class imbalance, mean validation loss and F1-score will be used to select final configuration as opposed to accuracy alone in line with other experts' recommendations [23]. The maximum epochs for training the network will be 200 epochs, and an early stopping using validation loss will be applied with a patience of 20 epochs. If validation performance will fail to improve, training will be terminated and the weights corresponding to the best performance restored.

The following is the update rule for weights (W):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

where:

- The learning rate, or η, is a hyperparameter that regulates the size of the weight update steps.
- $\frac{\partial \text{Loss}}{\partial W}$ is the gradient of the loss function in relation to the weights.

3.5.3. Overfitting Control and Dimensonality Reduction

To overcome the challenge of overfitting during model development as discussed, we will adopt dimensionality reduction, k-fold cross-validation, an independent 30% hold-out test set, early stopping and restriction of network complexity as confirmed by Khan and others [32]. In the event of class imbalance observation in the training data, SMOTE will be applied only to the training folds, thus avoiding information leakage into validation or test observations. Either, assessing model generalization will be done by comparing training, cross-validation, and independent-test performance. A substantial divergence between training and validation performance will be treated as evidence of overfitting, consistent with the backpropagation-based models' drawbacks as identified [20].

To minimize the number of features while maintaining the highest level of data variance Principal Component Analysis (PCA) as a dimensionality-reduction method will be applied. PCA is commonly used to achieve low-dimensional representation in machine learning [33]. PCA maps high-dimensional vector to a low-dimensional subspace representation through linear projection thereby aiding data visualization. Features comprising high correlation are dropped thus minimizing model instability as described in algorithm 1 below (Equation 2). Typically, a threshold of "0.9" is applied, and characteristics that exhibit a higher correlation than the designated threshold are eliminated. When working with complicated data, particularly high-dimensional feature sets, we shall employ t-SNE to visualize and reduce the dimensionality of the data.

Algorithm 1: Algorithm for dimensionality reduction

Algorithm 1: Drop correlated features from the dataset (2)

Input: Benign Data 2D Array, N, Correlation Threshold

Output: Benign Data 2D Array, Dropped Columns

1: correlation_matrix ← data.corr().abs()

2: upper_matrix ← correlation_matrix [i, j] {[i, j] ∈ N: i <= j}

3: dropped ← i { i ∈ N**:** correlation_matrix [i,*] > threshold}

4: data ← data.drop_columns (dropped)

5: **return** data, dropped

3.6. Data Analysis and Presentation

The data collected was subjected to statistical analysis using SPSS version 27 to draw research suppositions. Both descriptive and inferential statistics were required. The hypothesis test was specifically applied using chi-square. Qualitative data analysis provided rich, contextual insights that complemented quantitative data for the prediction of infectious diseases [16, 33]. We systematically analyzed qualitative data, identified relevant themes, and converted them into machine-readable formats. This provided useful insights to enhance the predictive power and diagnostic accuracy of the proposed model at later stages. To present the analyzed data, tables, charts, and graphs are used.

3.7. Ethical Considerations

The researchers ensured informed consent, and the privacy of the participants. Ethical approval was acquired from the relevant authorities, including the Board of Postgraduate Studies (BPS) of Murang'a University of Technology. A research permit was also obtained from the National Commission for Science, Technology, and Innovation (NACOSTI). All Personal Identifiable Information (PII) such as names, phone numbers, or national IDs were removed or anonymized before storage. We clearly explained the purpose of the study, how privacy was protected, and the potential benefits to the participants. They were made aware of their voluntary participation and their rights to excuse themselves at any level during participation. All materials or ideas used in this study were dully cited and we ensured the articles used were peer reviewed and published on reputable journals.

4. RESULTS

This section presents the results obtained from the field as analyzed.

4.1. Demographic Characteristics of Respondents

The demographic characteristics of the respondents were analyzed using frequencies and percentages with the aim of establishing the socioeconomic background of the study population. Specific variables examined included age group, gender, education level, occupation, household size, housing type, and residential area. The findings indicated that the majority of the respondents at 45.7% were aged 18-30 years, followed by those aged 31-45 at 23.9%, aged 46-60 at 15.2%, and aged 61 and above at 8.7%. Those under 18 years were 6.5%. Generally, the majority of the respondents were aged 18 to 45 years, which reflects the active working-age population within the informal settlements.

Female respondents were slightly more than male respondents, accounting for 52.2% and 47.8% respectively, indicating a relatively balanced gender representation. Most respondents had attained either primary education (39.1%) or secondary education (32.6%) while a smaller portion had college or University education (15.2%). Those without any formal education were 13%.

The occupational distribution revealed that a large proportion of respondents were engaged in informal economic activities such as casual labour, small-scale trading, and construction work, while analysis of the housing conditions revealed that most respondents resided in semi-permanent and temporary structures. Either household sizes indicated that majority at 39.1% were living in overcrowded households with more than five occupants, with 3-5 occupants at 34.8%, and 1-2 occupants being the minority at 26.1%.

4.2. Infectious Disease Patterns in Informal Settlements

The study sought to determine the distribution of infectious diseases in the informal settlements selected. Respondents were asked if they had ever been diagnosed with Tuberculosis, Malaria, HIV/IADs or Cholera, and their responses were captured. The results indicated a high prevalence of tuberculosis among the respondents, accounting for 34.8% of confirmed self-reported cases. Malaria followed closely at 30.4%, Cholera at 15.2%, while HIV/AIDs recorded the lowest prevalence at 8.7% of the cases. 10.9% of the respondents reported no confirmed cases.

4.3. Environmental Risk Factors

The study also sought to understand the environmental risk factors associated with infectious diseases in informal settlements. Frequency analysis, crosstabulations, and Chi-square tests were conducted to determine the relationship between environmental exposures and infectious disease occurrence among respondents. The correlation test was conducted at 5% level of significance with a 2-tailed test. Thus, the significance critical value was set at 0.025 above which the association is deemed to be insignificant and vice versa.

4.3.1. Tuberculosis Environmental Risk Factors

The findings revealed a significant correlation between TB and exposure to smoke, dust, and direct contact with infected persons, overcrowded environments, and smoking history. Respondents exposed to dust recorded considerably higher TB prevalence than their unexposed counterparts. Exposure to dust particularly showed a statistically significant correlation with tuberculosis infection ($\chi^2 = 12.530$, $p < .001$).

A similar analysis on the relationship between exposure to smoke and TB infection further demonstrated a strong association ($\chi^2 = 26.764$, $p < .001$), indicating that both indoor pollution and frequent smoke inhalation significantly increase risks of respiratory infection. The study also provided strong links between direct contact with sick persons with TB prevalence ($\chi^2 = 14.186$, $p < .001$), and likewise, with overcrowded environments, which also demonstrated a very strong correlation with TB infection ($\chi^2 = 14.382$, $p < .001$).

4.3.2. Malaria Risk Factors

The study demonstrated a strong relationship between malaria infection and mosquito-related environmental conditions. Frequent mosquito exposure indicated statistically significant relationship with malaria infection ($\chi^2 = 14.162$, $p < .05$). Stagnant water near residential areas additionally showed a highly significant association with malaria prevalence ($\chi^2 = 26.162$, $p < .05$). The presence of mosquitoes inside residential houses also recorded a very strong association with malaria infection ($\chi^2 = 46.000$, $p < .05$).

4.3.3. Cholera Risk Factors

The study also sought to establish sanitation-related risk factors associated with Cholera prevalence. The findings revealed that exposure to sewage have a statistically significant association with Cholera infection ($\chi^2 = 13.443$, $p < .001$) as well as lack of proper toilet facilities ($\chi^2 = 18.872$, $p < .001$). Irregular garbage collection significantly influenced Cholera prevalence ($\chi^2 = 10.733$, $p = 0.013$), whereas contact with dirty water showed moderate but important association with Cholera infection.

4.4. Symptoms Distribution

Respondents were asked if they had experienced some of the symptoms associated with tuberculosis, malaria, cholera, and HIV/AIDs. The study analyzed the distribution of these symptoms and the findings showed that fever was the most commonly reported symptom, affecting 60.9% of the respondents. Persistent cough accounted for 41.3%, while night sweats was 34.8%. Severe headache, vomiting, diarrhoea, and drastic weight loss reported varying proportions of respondents. Statistical analysis was conducted to examine the relationship between symptoms and diseases, whereby persistent cough and night sweats were found to be strongly associated with Tuberculosis, whereas fever and severe headache were primarily associated with malaria infection. Diarrhoea and vomiting were strongly linked to cholera infection. Weight loss demonstrated a statistically significant relationship with HIV/AIDs infection.

4.5. Existing Prevention and Management Strategies for Infectious Diseases

4.5.1. Prevention Practices

The study sought to understand the existing prevention and management strategies adopted by residents within the study areas. Descriptive statistical analyses were used to determine the extent to which respondents practiced preventive measures against the infectious diseases. The results indicated that HIV/AIDs prevention practices such as the use of condoms recorded the highest mean score (Mean = 4.13, SD = 0.980), suggesting relatively high awareness and adherence to HIV prevention interventions.

Handwashing with soap as a strategy against cholera also recorded relatively high mean score (Mean = 3.85, SD = 1.053), indicating moderate adoption of hygiene practices among respondents. Drinking treated water (Mean = 3.63, SD = 1.040), also demonstrated a moderate implementation as a prevention tactic against cholera. Sleeping under mosquito net (Mean = 3.61, SD = 1.201) to prevent malaria was also moderately practiced. Opening windows for ventilation however, recorded a comparatively lower mean score (Mean = 3.46, SD = 1.206), suggesting inconsistent ventilation practices despite the high prevalence of respiratory infections.

Table 1. Prevention Practices

Prevention Strategy	Mean (M)	Standard Deviation (SD)	Interpretation
HIV Prevention Practices	4.13	0.980	Highly Practiced
Wash Hands with Soap	3.85	1.053	Moderately Practiced
Drink Treated Water	3.63	1.040	Moderately Practiced
Sleep Under Mosquito Net	3.61	1.201	Moderately Practiced
Seek Medical Care When Sick	3.61	1.043	Moderately Practiced
Open Windows for Ventilation	3.46	1.206	Moderately Practiced

4.5.2. Health Seeking Behavior

The healthcare-seeking behavior among the respondents in relation to infectious disease management was examined. The findings revealed that 95.7% of the respondents sought treatment whenever symptoms appeared revealing relatively high utilization of healthcare within the study areas. Only 34.9% had tested for infectious diseases under study in the last one year. Pharmacies emerged as the most preferred first healthcare option accounting for 50.0% of responses. Public hospitals and private clinics followed at 23.9% and 19.6% respectively. Self-medication and traditional medicine options accounted for 4.3% and 2.2% respectively.

The descriptive analysis further showed that respondents demonstrated comparatively high health-seeking behavior when sick (Mean = 3.61, SD = 1.043), and the majority of respondents (69.9%) reported adhering to prescribed

treatment regimens, indicating relatively high treatment adherence within the study population.

4.5.3. Healthcare Accessibility Challenges

Respondents were asked about challenges affecting their accessibility to healthcare in managing infectious diseases, and their responses using Likert scale and analyzed. The findings indicated that healthcare facilities were generally accessible geographically as shown by a high mean score for ease of reaching healthcare facilities (Mean = 4.04, SD = 0.759). However, affordability of healthcare services remained relatively low (Mean = 2.74, SD = 0.743), suggesting that financial barriers continue to hinder access to effective treatment in study areas.

Fear of HIV and TB related stigma recorded a moderate-to-high mean score (Mean = 3.70, SD = 1.428), indicating that stigma remains a significant social challenge affecting healthcare utilization within slums. Availability of drugs at healthcare facilities presented a moderate mean score (Mean = 3.09, SD = 0.725), suggesting occasional shortage of essential medications.

Table 2. Healthcare Accessibilty challenges

Variable	Mean (*M*)	Standard Deviation (SD)	Interpretation
Health Facility Easy to Reach	4.04	0.759	High Accessibility
Fear of HIV/TB Stigma	3.70	1.428	Significant Challenge
Drugs Available at Clinic	3.09	0.725	Moderate Availability
Health Services Affordable	2.74	0.743	Low Affordability

4.6. Association Between Risk Factors and Infectious Diseases

To establish relationships between environmental risk factors and infectious diseases under study, Spearman correlation analysis and Chi-square tests were conducted. The findings revealed a strong positive correlation between exposure to crowded environments and tuberculosis prevalence (r=0.559, $p < .001$). Exposure to smoke also demonstrated a strong positive relationship with tuberculosis infection ($r = 0.630$, $p < .001$). Malaria prevalence demonstrated strong positive correlation with stagnant water near residential areas ($r = 0.583$, $p < .001$), and the presence of mosquitoes inside houses ($r = 0.827$, $p < .05$). Cholera prevalence showed a strong positive relationship with exposure to sewage ($r = 0.541$, $p < .001$) and exposure to sewage ($r = 0.437$, $p = 0.002$). HIV/AIDs prevalence indicated a statistically significant relationship with pre-existing medical conditions ($r = 0.467$, $p = 0.001$).

Table 3. Association between risk factors and infectious diseases

Risk Factor	Disease	Correlation Coefficient (*r*)	*p*-value	Interpretation
Exposure to crowded environments	Tuberculosis	0.559	0.000	Strong positive
Exposure to smoke	Tuberculosis	0.630	0.000	Strong Positive
Frequent mosquito exposure	Malaria	0.555	0.000	Strong Positive
Stagnant water near house	Malaria	0.683	0.000	Strong Positive
Mosquitoes common inside house	Malaria	0.827	0.000	Very Strong Positive
Exposure to sewage	Cholera	0.541	0.000	Strong Positive
Stagnant water near house	Cholera	0.437	0.002	Moderate Positive

5. DISCUSSION

The study established that Tuberculosis was the most prevalent infectious disease among the respondents in Kibera, Mathare, and Mukuru kwa Njenga slums, followed closely by Malaria. Cholera and HIV/AIDs followed at distant third and fourth respectively. Tuberculosis prevalence was strongly

associated with overcrowding, exposure to smoke, exposure to dust, smoking history, and direct contact with infected persons which corroborates the findings of similar studies [1, 4]. Malaria prevalence was significantly influenced by mosquito exposure, stagnant water accumulation near houses, and the presence of mosquitoes inside residential buildings. Cholera prevalence was linked to poor sanitation conditions, exposure to sewage, inadequate toilet facilities, and irregular garbage collection systems.

Regarding prevention and management strategies, we established that respondents demonstrated moderate implementation against infectious diseases. Nonetheless, affordability of healthcare services, stigma associated with HIV and TB, and inconsistent drug availability remained significant barriers to effective disease management. These findings further illustrate why TB remains notably prevalent as compared to other diseases.

The correlation and Chi-Square analyses conducted confirmed statistically significant relationship between environmental risk factors and infectious disease prevalence, thereby supporting the study hypothesis that environmental and socioeconomic conditions within informal settlements significantly influence infectious disease transmission.

5.1. Association Between Study Outcome and Model Development

The findings of this study in an attempt to address Objective (i), outlines the empirical basis for determining the epidemiological relevance of the variables in the training dataset to be selected, and the study settings. Accordingly, the findings inform feature interpretation and contextual validity, whereas the external large dataset to be selected at a later stage, will support model training and technical validation. We propose subsequent external validation of the resulting model using TB records from Kenyan informal settlements before claims of local clinical applicability are made.

5.2. Integration with the Backpropagation Algorithm

The analysis in this paper provides critical grounds for the achievement of subsequent objectives. Based on the outcome which ranks Tuberculosis as the most prevalent infection among the respondents, this study picks TB as the infectious disease of choice for training the proposed model, making it both disease and setting specific. The statistically significant variables identified through Chi-square, correlation, and frequency analyses in this section will be important in identifying input variables from the training dataset for the model. In particular, demographic, clinical-history, HIV-status, specimen-type and TB-related clinical information are some of the possible input variables that will be considered while TB prediction will serve as the output variable. As observed, the findings of this paper provide empirical evidence of key risk factors and form the basis for the selection of input variables for the backpropagation neural network model developed in the subsequent sections.

All predictor variables that will be identified in the selected secondary dataset will be transformed into numeric formats in readiness to integration with the model training. Binary variables will be maintained in 0/1 format, while ordinal demographic variables will be encoded as integer values. This transformation will ensure that the raining dataset meet the input requirements for artificial neural networks, which require numeric feature vectors. The backpropagation neural network will be trained to predict infectious disease occurrence based on the risk factors already identified. This section aligns with our recommendations for disease predictive models to be grounded in empirical data that reflect lived conditions in targeted settings to minimize misclassification and equity aspects.

6. CONCLUSIONS AND RECOMMENDATIONS

This paper confirms high risks of infectious diseases in informal settlement with TB being the highest among the respondents. It highlights challenges with existing management strategies and lays grounds for the development of a setting-specific sustainable machine learning model to boost prediction within low-resource environments.

Cognizant of its limitations, we recommend implementing and validating the model using real-world datasets collected from informal settlements across the world to determine its predictive performance under varied practical conditions. Future research should evaluate the comparative performance of stochastic, batch, and adaptive optimization techniques, investigate the integration of multimodal datasets such as environmental and geospatial information, and examine model explainability to improve trust and adoption. Either, governments, healthcare institutions, and development partners should invest in digital health infrastructure, standardized health data collection, and interoperable health information systems that facilitate continuous model learning and deployment.

6.1. Study Implications

This research contributes to the growing application of artificial intelligence in public health by demonstrating how optimized backpropagation-based neural networks can support sustainable infectious disease prediction in resource-constrained environments. Its findings are expected to inform healthcare practitioners, policymakers, and researchers on the design of scalable, cost-effective predictive systems tailored to informal settlements. The paper also provides a foundation for integrating machine learning into disease surveillance, improving early diagnosis, optimizing healthcare resource allocation, and guiding future research on AI-driven public health interventions in developing countries.

7. ACKNOWLEDGEMENTS

We would like to thank the Journal’s editors and the reviewers for their valuable contribution in improving the quality and clarity of this manuscript.

8. REFERENCES

[1] Gulis, G., Mulumba, J. A. A., Juma, O., and Kakosova, B. 2014. Health status of people of slums in Nairobi, Kenya. *Environmental Research*, *96*(2), 219–227. https://doi.org/10.1016/j.envres.2004.01.016

[2] Shimpi, P., Shah, S., Shroff, M., and Godbole, A. 2017. An artificial neural network approach for classification of vector-borne diseases. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 412-415). IEEE. https://doi.org/10.1109/ICCMC.2017.8282721

[3] Bhimbhai, P. P., and Jagannath, P. N. 2022. Review on Deep Learning Techniques for the Diagnosis of Corona Virus Using CT & X-Ray Images. *Proceedings of the Advancement in Electronics & Communication Engineering*. https://dx.doi.org/10.2139/ssrn.4159103

[4] Tadesse, G. A. 2020. Multi-modal diagnosis of infectious diseases in the developing world. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 2131–2141. https://doi.org/10.1109/JBHI.2019.2950608

[5] Wilson, E. A. 2018. 'Don't say "research"': reducing bidirectional risk in Kibera slum. *Contemporary Social Science*, *13*(3-4), 337-353. https://doi.org/10.1080/21582041.2017.1418525

[6] Akatch, S. O., and Kasuku, S. O. 2019. Informal settlements and the role of infrastructure: The case of Kibera, Kenya. *Discovery and Innovation*, *14*(1–2), 32–37. https://doi.org/10.4314/dai.v14i1.15421

[7] Bhalla, J. J. S., and Aggarwal, A. 2013. A novel method for medical disease diagnosis using artificial neural networks based on backpropagation algorithm. In *2013 Confluence: The Next Generation Information Technology Summit (4th International Conference)* (pp. 55–61). IEEE. https://doi.org/10.1049/cp.2013.2293

[8] Burton, D. C., Flannery, B., Onyango, B., Larson, C., Alaii, J., Zhang, X., ... and Feikin, D. R. 2011. Healthcare-seeking behaviour for common infectious disease-related illnesses in rural Kenya: a community-based house-to-house survey. *Journal of health, population, and nutrition*, *29*(1), 61. https://doi.org/10.3329/jhpn.v29i1.7567

[9] Agatonovic-Kustrin, S., and Beresford, R. 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

[10] Showkatian, E., Salehi, M., Ghaffari, H., Reiazi, R., and Sadighi, N. 2022. Deep learning-based automatic detection of tuberculosis disease in chest X-ray images. *Polish journal of radiology*, *87*(1), 118-124. https://doi.org/10.5114/pjr.2022.113435

[11] Kaur, I., Sandhu, A. K., and Kumar, Y. 2022. Artificial intelligence techniques for predictive modeling of vector-borne diseases and its pathogens: a systematic review. *Archives of Computational Methods in Engineering*, *29*(6), 3741-3771. https://doi.org/10.1007/s11831-022-09724-9

[12] Fathima, S., and Hundewale, N. 2011. Comparison of classification techniques—SVM and Naive Bayes to predict the arboviral disease dengue. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)* (pp. 538–539). IEEE. https://doi.org/10.1109/BIBMW.2011.6112426

[13] Lee, M. K., Paik, J. H., and Na, I. S. 2020. Outbreak prediction of hepatitis A in Korea based on statistical analysis and LSTM network. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 379-381). IEEE. https://doi.org/10.1109/ICAIIC48513.2020.9065082

[14] Jahnavi, K. 2023. *Detection of COVID-19 using ResNet50, VGG19, MobileNet, and forecasting using logistic regression, Prophet, and SEIRD model*. https://doi.org/10.48550/arXiv.2304.07617

[15] Subiyanto, Mulwinda, A., and Andriani, D. 2017. Intelligent diagnosis system for acute respiratory infection in infants. In *2017 3rd International Conference on Science in Information Technology (ICSITech)* (pp. 558–562). IEEE. https://doi.org/10.1109/ICSITech.2017.8257175

[16] Gibson, N., Vertigans, S., and Mueller-Hirth, N. 2023. Empowering Kibera during political change: A case study through a Freirean lens. *International social work*, *66*(2), 388-404. https://doi.org/10.1177/00208728211010215

[17] Pathmind. (n.d.). *A beginner's guide to backpropagation in neural networks*. https://wiki.pathmind.com/backpropagation

[18] Tadesse, G. A., Zhu, T., Le Nguyen Thanh, N., Hung, N. T., Duong, H. T. H., Khanh, T. H., ... and Clifton, D. A. 2020. Severity detection tool for patients with infectious disease. *Healthcare Technology Letters*, *7*(2), 45-50. https://doi.org/10.1049/htl.2019.0030

[19] Nicholson, C. 2020. *A Beginner's Guide to Backpropagation in Neural Networks*.

[20] TensorFlow. 2020. *TensorFlow API documentation*. https://www.tensorflow.org/api_docs/python/tf

[21] Astuti, I., and Sutarno, H. 2017. The expert system of children's digestive tract diseases diagnostic using combination of forward chaining and certainty factor methods. In *2017 3rd International Conference on Science in Information Technology (ICSITech)* (pp. 608-612). IEEE. https://doi.org/10.1109/ICSITech.2017.8257185

[22] Kayali, I. 2018. Expert system for diagnosis of chest diseases using neural networks. *arXiv preprint arXiv:1802.06866*. https://doi.org/10.48550/arXiv.1802.06866

[23] Chicco, D., and Jurman, G. 2021. An ensemble learning approach for enhanced classification of patients with

hepatitis and cirrhosis. *IEEE Access*, *9*, 24485–24498. https://doi.org/10.1109/ACCESS.2021.3057192

[24] Di Martino, F., Delmastro, F., and Dolciotti, C. 2021. Malnutrition risk assessment in frail older adults using m-health and machine learning. In *ICC 2021-IEEE International Conference on Communications* (pp. 1-6). IEEE. https://doi.org/10.1109/ICC42927.2021.9500471

[25] Morens, D. M., & Fauci, A. S. 2013. Emerging infectious diseases: threats to human health and global stability. *PLoS pathogens*, *9*(7), e1003467. https://doi.org/10.1371/journal.ppat.1003467

[26] Hossain, S. I., Nipu, S. A., and Hasan, M. R. 2023. Recognition of tuberculosis on medical x-ray images utilizing mobilenet transfer learning. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (pp. 952-959). IEEE. https://doi.org/10.1109/ICAIS56108.2023.10073728

[27] Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., and Napolitano, A. 2012. A review of the stability of feature selection techniques for bioinformatics data. In 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI) (pp. 356–363). IEEE. https://doi.org/10.1109/IRI.2012.6303031

[28] Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., ... and Das, R. 2016. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR medical informatics*, *4*(3), e5909. https://doi.org/10.2196/medinform.5909

[29] Adeyemo, O., Adeyeye, T., and Ogunbiyi, D. 2015. Comparative study of ID3/C4.5 decision tree and multilayer perceptron algorithms for the prediction of typhoid fever. *African Journal of Computing & ICT*, *8*(1), 185–192.

[30] Delahunt, C. B., Mehanian, C., Hu, L., McGuire, S. K., Champlin, C. R., Horning, M. P., ... and Thompon, C. M. 2015. Automated microscopy and machine learning for expert-level malaria field diagnosis. In *2015 IEEE global humanitarian technology conference (GHTC)* (pp. 393-399). IEEE. https://doi.org/10.1109/GHTC.2015.7344002

[31] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77. https://doi.org/10.2753/MIS0742-1222240302

[32] Khan, M. T., Kaushik, A. C., Ji, L., Malik, S. I., Ali, S., and Wei, D. Q. 2019. Artificial neural networks for prediction of tuberculosis disease. *Frontiers in microbiology*, *10*, 395. https://doi.org/10.3389/fmicb.2019.00395

[33] Park, S., and Cho, E. 2014. National infectious diseases surveillance data of South Korea. *Epidemiology and health*, *36*, e2014030. https://doi.org/10.4178/epih/e2014030

[34] Bhatt, T. 2020. Introduction to gradient descent algorithm, along with its variants. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/introduction-to-gradient-descent-algorithm-along-its-variants/