Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Formulation of a Multi-Disease Comorbidity Prediction Framework: A Data-Driven Case Analysis on of Diabetes, Hypertension, and Cardiovascular Risk Trajectories

Md Zakir Hossain[1] ✉ Md Munsur Khan[2], Raqibul Islam[3], Kamrun Nahar[4] and Md Firoz Kabir[5]
[1]Mathematics, Statistics & Computer Science Department, University of Wisconsin-Stout, Menomonie, WI, USA
[2]Department of Computer Science and Engineering, Eastern University, Dhaka, Bangladesh
[34]School of Management, Kettering University, Flint, MI, USA
[5]Manager (IT Dep.), Language Help bd, Bangladesh
Corresponding Author: Md Zakir Hossain, **E-mail**: mhossain8@my.gcu.edu

| ABSTRACT

The development of a multi-disease comorbidity prediction framework is the main drive point of this study, leading to the data-driven case analysis of diabetes, hypertension, and cardiovascular risk trajectories. As non-communicable diseases are showing more and more overlapping patterns, the study is able to predict how to create that model by taking into consideration the co-occurrence instead of treating diseases in a separate way. Using a well-organized medical dataset that holds demographic and clinical factors that include age, gender, BMI, smoking history, blood glucose, and HbA1c levels, the study uses linear regression methods to evaluate risk factors of individual and cumulative disease impacts. The deployed methodology intertwines the strong preprocessing procedures, correlation analysis, and training of the model, and final overall visualization. The data visualization tools like Python, Tableau, and Microsoft Excel may also be employed to assist in obtaining insights, revealing the trends, and presenting the findings. The major trends diagnosed are the increased risk of co-morbidity in aged people, increases in BMI and level of HbA1c, the exacerbating effect of a smoking history. These results point to the interdependence of metabolic health and cardiovascular health and indicate the need of advanced predictive pathways in clinical practice. This study also shows that visual analytics in tandem with explanatory statistics modeling can be very relevant when diagnosing cases at the early stage and preventive health measures. This study is meaningful to producing a feasible and translatable methodology of precision public health by placing an emphasis on establishing transparency, scalability, and ease of implementation. The suggested model is a kind of decision aid which can help medical workers of the sphere to actively diagnose people at risk and provide specific interventions even before the disease could become difficult to treat or impossible to reverse.

| KEYWORDS

Comorbidity Prediction, Machine Learning in Healthcare, Heart Disease and Hypertension Risk, Diabetes Forecasting, Multi-Label classification and Chronic Disease Analytics.

| ARTICLE INFORMATION

ACCEPTED: 02 July 2023 **PUBLISHED:** 25 July 2023 **DOI:** 10.32996/jcsts.2023.5.3.12

1. Introduction

1.1 Background

Chronic illnesses like diabetes, high blood pressure and cardiovascular diseases (CVD) have been on the rise as the primary contributors to morbidity and death. Nowadays, it is no longer isolated conditions as these chronic conditions co-occur in patients, establishing interdependencies that are difficult to deal with using traditional models in healthcare. As digital health technologies and data have become available, the approach of reactive treatments can be changed to proactive risk management and diagnosis of such diseases at a very early stage. Diabetes is one of them: as a condition in itself, it is a serious threat in its own right, and as a precondition to the even more dangerous diseases like hypertension and heart disease. The World Health Organization estimated that in the year 2021, more than 537 million adults harbor diabetes, and the number is likely to increase tremendously. As well, ill patients, who have been diagnosed with diabetes conditions, have increased risk of developing hypertension and heart disorders [1]. The common trajectories imply the need of a broader analytical methodology which has to go beyond standalone predictors of disease and extend the existing risk in multi-disease context. Current machine learning methods hold novel prospects in terms of the study of various data types and identification of unobvious, non-linear dependencies between medical parameters. Algorithms presented in the form of patient data: including age, gender, BMI, smoking history, and clinical biomarkers, are capable of unearthing risk patterns that are not readily discernible by traditional means. The research is based on the exploitation of an organized table with meaningful health indicators to construct a combined predictive framework. In this way, the study would be of benefit as the disease prevention and control methods would be easy to interpret and scale, which, in turn, would help increase the quality of healthcare decision-making and individual treatment planning.

1.2 Innovations in Technology with regards to Predictive Healthcare

The dynamically increasing amount of health-related information and the development of artificial intelligence (AI) changed the prospects of predictive health significantly. The increasing amounts of the electronic form of medical record, biometrics and population health data allow opportunities to be accessed to derive useful information to prevent and manage diseases. A subfield of AI, such as machine learning, has been specifically promising in the fields of healthcare because of its capacity to discern latent patterns, non-linear connections, and complex sets of relationships between factors. Research on algorithms like logistic regression, decision trees, support vector machines, and ensemble models like XGBoost and Random Forest is aggressively taking place to diagnose this in the early stages, segment the patients and stratify the risk among patients. A single-disease form of classification, i.e., the input-output learning/classification of a single disease, e.g., predicting diabetes or cardiovascular disease alone, has been most popular in the machine learning featured in healthcare. This is a pitfall to the provision of holistic care since many patients diagnosed with one chronic condition acquire or develop other ones, like hypertension and heart disease [2].. This gap has a chance to be altered with the appearance of multi-label classification techniques. With the help of multi-label solutions, one can develop a co-morbidity predictor that could estimate the current co-occurrence or potential occurrence of several chronic ailments. Further, explainable AI tools can be applied, like SHAPE and make such models more transparent and trustworthy, finding out which variables are the most important to guess outcomes [3]. The application of such a strong, evidence-based framework that incorporates all these tools may therefore be central in ensuring smarter healthcare solutions are achieved at a patient level, as this study suggests.

1.3 Problem Statement

Although there have been increasing developments in the field of predictive healthcare, the majority of the existing models are single-disease based models that do not take into consideration the dynamics of a co-morbid condition. The problem of such a multi-disease prediction is often the development of hypertension and cardiovascular disease in patients with diabetes, but few studies focus on the complex problem [4]. The absence of integrated models lowers the efficiency of early interventions and individual care. Thus, it is high time to develop a data-driven framework of co-morbidity prediction that will take into account a variety of conditions and determine the interdependence of risk factors within the demographic and medical profile of a patient.

1.4 Research Objectives

In this study, the authors want to propose an integrated machine learning pipeline based framework that can predict all the three comorbidities, diabetes, hypertension and cardiovascular, simultaneously.

- To investigate real-world medical data with demographic and biometric indicators and pre-process it.
- To develop binary and multi-label classification models of predicting the availability of several diseases.
- With the help of the interpretation tools like SHAP to find out the most powerful things causing comorbidity development.
- To determine the predictive effect and the robustness of machine learning algorithms in terms of multi-disease risk prediction.
- To profile patients who have comorbid conditions in order to determine the course of comorbidity.
- In order to suggest a clinical application model capable of facilitating the early risk detection and focused intervention in healthcare facilities.

1.5 Theoretical Framework

This study is theoretical in nature using a number of interconnecting theories that influence the direction and inference of multi-disease prediction. Fundamentally, the study is based on an Epidemiologic Triangle concept, sometimes referred to as ETA, comprising host, agent, and environmental factors that explain the diseases via the interactions among the three categories of risk aspects, namely individual, biological, and environmental risk. Here, age, BMI, smoking background, and glucose will be variables that can be used as a measure of an individual's vulnerability and exposure [5]. The Health Belief Model (HBM) argues that the risk perception, when informed with predictive models, might result in improved health seeking behaviors interventions. In the technological dimension, the Data-Driven Decision-Making Framework provides the basis of using empirical evidence in reinforcing clinical decisions. This model assumes that incorporating predictive analytics in the healthcare process may be capable of converting raw patient data into actionable information. The methodological framework of this study is offered according to the Machine Learning Pipeline Theory, i.e., the data preprocessing and features selection stages, the training, validation, and interpretability procedures of the models. The transparency achieved through SHAP value is also consistent with the Explainable AI (XAI) Paradigm that focuses on accountability and fairness of all models in such critical applications as healthcare [6]. The combination of these theories will help to merge the theoretical modeling and practical use of digital health, which is the purpose of the proposed study. The combination between epidemiological reasoning and state-of-the-art machine learning allows a ubiquitous interpretation of the co-occurrence of disease and maximizes the viability of introducing personalized, preventive, and scalable measures in a clinical environment.

1.6 Research Questions

The proposed investigation aims at answering the following major questions concerning the multi-disease prediction through the data analytics and machine learning models in the clinical practice.

1. To what extent could structure patient data be used for predictions of co-occurring diseases like diabetes, hypertension, heart disease using machine learning models?
2. Which are the most relevant demographic and medical characteristics that increase the probability of acquiring such a number of chronic diseases at the same time?
3. What would a data-driven framework help to contribute to early detection and clinical decision-making, preventive care of patients with the risk of chronic comorbidities?

1.7 Significance of the Study

This study has a huge implication on the scientific community and the general health space worldwide. With the emphasis on co-morbidities, namely, diabetes, hypertension, and cardiovascular diseases, the research concerns a topical global issue that worries old and aging populations and low- and middle-income countries in particular. Conventional healthcare systems usually address such illnesses separately, without relying on the fact that such diseases are bound together in relation to one another and have an added threat towards the concerned patients. Not only is a multi-disease prediction model more accurate in terms of diagnosing but also it is the basis of proactive healthcare. Machine learning implementation in the study of disease co-occurrence can be used to come up with intelligent clinical decision support systems (CDSS). Such systems would help healthcare providers to recognize those who are high-risk, offer an individualized plan of treatment, and enhance patient monitoring [7]. In preventive programs as well, the resources may be distributed more efficiently and outreach can be focused more precisely thanks to predictive modeling. The work is of greater importance to the sphere of digital health and precision medicine. Explainable AI being embedded with the SHAPE model also makes clinical tools transparent thus enabling a user to know why his or her prediction is as such. Such interpretability is essential with regard to providing trust and adoption to the real world [9]. This study will enable groundbreaking transfer between raw health data and meaningful action, paving the way to data-informed public health methods that are affordable, scalable, and effective.

1.8 Scope and Delimitations

This study research is aimed at developing and analyzing a machine learning framework that can predict co-morbidities namely diabetes, hypertension or cardiovascular diseases. Data included in the study is structured and tabular data containing a combination of two kinds, demographic and clinical indicators such as age, gender, smoking history, and HbA1c, BMI, blood glucose. The models will be restricted to the supervised learning methods, Logistic Regression, Random Forest, and XGBoost compared with their performance. The study is concerned more with binary and multi-label classification, or in other words, it is related to classification with the aim of identifying whether these conditions are evidence or not, but not whether they are at an early or a late stage of the condition [10]. Dynamic variables medical data are not considered; the framework does not allow modeling the progression of the disease with time. Besides, the data does not include detailed patient history, history or medicines used or genetic-related data, which would increase the predictive capability. It is not the aim of the research to substitute medical diagnostics but it is aimed to complement it with the guidance that can be made based on a recently identified pattern of risks. The issue of data privacy, model transparency, and bias in predictions are considered as ethical, whereas implementing the regulatory compliance frameworks like HIPAA or GDPR is beyond the scope of this research. Although these results are encouraging when it comes to digital interventions in health, they might not be applicable to other groups, especially individuals who have varying access to health or socio-environmental situations [11]. Further study may be done regarding cross-domain validation and model extension in order to consider more diseases or longitudinal medical history

2. Literature Review

2.1 Concept of Comorbidity Chronic Disease

Comorbidity between chronic diseases, mainly diabetes, hypertension and cardiovascular disease (CVD), is a major health issue to the global healthcare community. Having more than one chronic condition in a patient at the same time creates complications in diagnosis, treatment protracted care. This normally results in poor living standards, high mortality rate, and escalating medical costs. The similarity in these conditions is due to the excessive metabolic, behavioral, and lifestyle risk factors, including obesity, inappropriate diet, sedentary lifetime activities, and smoking [12]. The inseparable characteristic of these diseases has shown the problem of the traditional, soloed approach in medicine, which in most cases means that each disease is considered separately in its own context rather than validating the overall perspective of a patient. In health care practice, it results in incomplete care, duplication of diagnosis, and conflicting therapy. Amidst the rise in the prevalence of chronic conditions, there is an urgent demand to create integrative models and approaches by taking into consideration a combination of conditions. New opportunities have arisen with the recent advances in computational modeling and the availability of more data to construct predictive tools on a holistic basis. These tools can exploit structured data to predict the occurrence of several diseases at the same time, so that clinicians are able to treat the situation earlier and most efficiently [13]. With the rise of personalized medicine and use of data in decision making in healthcare, the use of predictive models that could integrate the patterns of co-morbidity becomes crucial. Understanding and being able to predict co-morbidity is important and in that way, the benefits people get are not just at personal levels but also within the health system in terms of improvement of planning, allocation of resources and policy. The purpose of this literature review is to discuss the extent to which the emerging machine learning techniques meet this acute healthcare priority.

2.2 Prediction models of diabetes

Diabetes prediction with machine learning models has been a rather popular direction since the prevalence of the disease has been growing significantly and there are properly structured medical datasets to work with. Such models have included logistic regression, decision trees, random forest and support vector machines where popular demographic and clinical characteristics such as age, BMI, blood glucose level, and HbA1c values have been used. These models are used to predict the chances of developing diabetes early so as to facilitate early diagnosis and preventive strategies. Most studies have only looked at diabetes as an individual phenomenon though they have been helpful in tasks related to binary classification. They fail to include co-existing conditions such as hypertension or cardiovascular disease, which often occurs together with diabetes in reality-based patients. This compartmentalized model largely works in solitary situations but is not reflective of clinical reality of multi-morbidity, thus not allowing the application of such models to work with the complex patient profiles. Besides, the majority of models are based on predictive accuracy without considering the transparency and explain ability of a model, which are fundamental factors of acceptance in clinical practice. Albeit the enhancement of the accuracy rates of diabetes prediction by exploiting ensemble learning and feature optimization, at the expense of interpretability this enhancement can be obtained. Clinicians might not trust or be willing to take an action based on predictions of opaque systems. Although the development of diabetes prediction is remarkable, the existing methods have to be updated so that they could be viewed through a wider prism of the diagnostic process and it is important to focus on interpretability that must be aligned better with the demands and challenges of contemporary healthcare delivery systems.

2.3 Modeling of Cardiovascular Risk and Hypertension

Heart disease, cardiovascular disease (CVD) and hypertension are some of the most common and lethal chronic diseases in the whole world. In regards to the diseases, predictive modeling has been almost exclusively on risk scores based on blood pressure, cholesterol, age, smoking status, and other clinical and lifestyle indicators. The use of traditional scoring systems in primary care utilizes an estimation of the 10-year risk of cardiovascular event, and these systems tend to fail in most aspects regarding adaptability and generalizability to other populations [14]. The development of machine learning techniques has led to the introduction of more dynamic and data-driven models that have better predictive powers with the capacity of integrating non-linear effects amongst the risk variables. These newer models use an expanded set of input factors, such as incorporating behavioral, environmental, and biometrics to develop a model of what the risk profile would be of individual patients. Even in the face of these advances, most systems risk models of cardiovascular and hypertension continue to coexist as they predict each condition in isolation. Such an approach fails to look at the further dynamics between CVD, hypertension, and other long-term illnesses like diabetes as they may commonly share some pathophysiological mechanisms. Numerous models applied in clinical practice are not explainable and therefore inapplicable [15]. Lack of built-in models that are able to be interrogated with reference to these intertwined health conditions is a huge gap in predictive medicine. Such bridging of the gap would not only enable more accurate stratification of disease risk but also would allow early intervention strategies that focus on patients at high risks of developing multiple chronic diseases at the same time.

2.4 Multi-Disease Prediction and Multi-Label Classification

The growing awareness of the co-occurrence of chronic diseases has had an influence on replacing the use of single-disease predictive models with more sophisticated, multi-disease frameworks. Multi-label classification (MLC) has become one of the useful tools in this regard, and it should be used to predict multiple medical outcomes of a single patient at a time. In contrast to regular binary classifiers, wherein application looks at the presence or absence of one condition, MLC models can take into consideration a complex comorbidity situation, i.e., complex conditions are prevalent, e.g., diabetes, hypertension, and cardiovascular disease. The latter based on structured datasets train algorithms to generate multiple disease labels on the basis of a shared set of input features, such as age, \(BMI\), blood glucose, and lifestyle variables [16]. The major strength of MLC is the ability to capture the interdependence and correlation between diseases, hence it makes the predictions more practical to clinical practice. Multi-label frameworks do not require multiple models to be executed in parallel, which further optimizes available computational resources, and eases the deployment within any healthcare environment. Nonetheless, the successful implementation of MLC in the healthcare system remains to be in its infancy. Most of the current implementations rely on open-ended and closed electronic health records (EHR) which are not available publicly hence reproducibility and scaling are constrained. Also, the interpretability of MLC models remains a challenge since most models are not interpretable at all [17]. The scope of these challenges can be resolved with the combination of the use of interpretable machine learning methods, through the utilization of more broadly applicable datasets. This will bring more application quality to multi-disease prediction models and will help transfer them to clinical operations.

2.5 Importance of Interpretability in Healthcare Models

At a time when machine learning has come to be the most vital element in the healthcare industry, interpretability has emerged as a key requirement in all applications of machine learning in healthcare. Although reaching high predictive accuracy is a technical challenge, clinicians and healthcare providers require transparency to trust and be able to put predictive models into a realistic practice. The traditional statistical models like logistic regression possesses interpretability as the clinicians can easily know the effect of the input variables on the final prediction [18]. These models can have an inability to profile multi-dimensional and non-linear data interrelation. Other, more complex models such as ensemble trees or deep neural networks are able to offer greater accuracy, but tend to have the limitation of being a black box providing little information on how decisions are reached. And the development of the explainable AI (XAI) frameworks overcome this trade-off by offering ways to explain more complicated models. Other methods such as SHAP values allow feature wise decomposition of predictions, so the most influencing variables become clear within a risk profile that the particular patient has. In such a level of granularity, healthcare providers can make informed decisions, support them and explain them adequately to patients on risks. In addition, interpretability increases the auditability of the model to ensure compliance requirements and ethical behaviors [19]. The interpretability in the multi-disease prediction context would also allow identifying the common risk factors, allowing more holistic and concerted treatment strategies. The availability of the interpretability tools should not be confined to a mere issue of technical strength, as it is an essential construct of clinical deployment, credibility, and, eventually, improved outcomes of patients.

2.6 Gapping of the research and justification.

In spite of the considerable advances in predictive healthcare, there exists one major gap in building interpretable multi-disease predictions models with accessible and well-structured data. Most existing examinations are short-term and are concerned about individual disease outcomes, and the use of sophisticated electronic health records, which can often be replicated, or inaccessible. Although such models are effective in a controlled setting, their applicability in a general case is limited and practically not applicable in areas with poor data infrastructure [20]. Most of the current frameworks do not provide provisions to co-morbid conditions without taking into account the interrelated nature of diseases such as diabetes, blood pressure, and heart problems. It is because of this lack that insights become fragmented and healthcare systems are simply unable to embrace truly integrative strategies. Most predictive models have been focused on the performance scores like accuracy, precision and recall measures, but lack adequate transparency and interpretability, which makes them less useful in clinical decision-making. There is a dire need and unmet demand to have a data-driven model of co-morbidity where the model can be understood. In this work, the researcher deals with this gap by using a structured, publicly available dataset with a significant number of demographic and clinical characteristics, using methods and models of multi-label classification and applying explainable AI. This kind of solution implies a sufficient predictive performance, being transparent and usable in actual work healthcare locations. Putting an emphasis on multi-disease prediction based on available data and understandable models, the study will advance the emerging area of personalized medicine / preventive medicine and enhance the concept of advanced healthcare analytics to be more inclusive, scalable, and clinically focused

2.7 Empirical Study

The article, “Dynamic Patient-Cluster Associations: Tracking Trajectories of Multiple Long-term Conditions” written by Ron Kremer and authorships are published at IEEE (2022), provides useful empirical observations regarding the multi-disease comorbidity modeling context. This study makes use of around 143,000 UK Biobank Electronic Health Records (EHRs) to study the dynamics of multi morbidity using time-independent clustering using Latent Dirichlet Allocation (LDA) which is a topic modeling procedure. Instead of making certain forecasts, the study provides a measure of a new metric that reflects the gravitational pull or affinity of one patient to the evolving clusters of diseases as time progresses. The dynamic model helps the researchers monitor co-morbid courses and set up essential windows of preventive interventions without the demand of complicated forecasting models which usually have confining information. This article is of utmost interest to the current research that also heavily relies on the comprehension of non-static, data-based perception of such diseases such as: diabetes, hypertension and cardiovascular risk [1]. The methodology developed by Kremer et al. allows arguing the hypothesis that co-occurrence among chronic diseases can be improved with the help of dynamic modeling and cluster evolution, instead of narrow binary forecasts. The fact that they established a data-driven basis in their research gives more credence to the argument that research should build on coherent explicable models that support personalized healthcare plans. It does the same concerning the crucial issues in the area of traditional predictive modeling, keeping the temporal integrity intact and being able to explore the patient health pathway in real-time, which directly corresponds to the goals of the presented study.

In the article by Ayush Jain et al., Generating New Drug Repurposing Hypotheses Using Disease-Specific Hypergraphs, Hayes introduces a strong computational framework of improving disease-specific modelling using disease-specific hypergraph and is a major weakness of the traditional modelling of drug repurposing using networks. The authors suggest an approach to building specific diseases hypergraphs instead of general intercoms by representing biological pathways as hyperedges and embedding them into a hypergraph with a modified node2vec algorithm. They have used their empirical assessment on Alzheimer disease to distinguish seven repurposing candidates neglected by the traditional models and this shows the significance of representing diseases using network architectures. Even though this technique is narrowly focused on the discovery of drugs, it is also a good analogy in co-morbidity modeling in healthcare analytics where specificity of disease and interaction patterns is paramount [2]. Disease-specific hypergraph concept can be generalized to the multi-disease contexts, e.g., diabetes, hypertension, cardiovascular diseases, to name just a few, in the context of determining both common biological processes and predictive risk patterns in multi-morbid populations, thus making the proposed study a beckoning empirical source of information in the context of this research.

The research findings within the dissertation titled as Artificial Intelligence Methods for Discovery in Large Biobanks (Lu Yang, 2023) are an elaborate info dump into the potential usage of cutting-edge AI techniques to conduct pattern discovery, disease prognostication, and phenotype extraction of massive biomedical datasets, such as the UK Biobank. Yang provides the scalable algorithms with usage of deep learning and probabilistic graphical models and surface latent relationships between genetic, phenotypic, and clinical variables. Another important aspect of this work is that it can process high-dimensional, multimodal data which is bringing together electronic health records (EHRs), genomics, imaging, and lifestyle factors, to determine co-occurring conditions and paths of disease progression. The given strategy directly fulfills the goals of multi-disease prediction research, as it allows organizing a population-scale cross-condition risk analysis [3]. The dissertation focuses on interpretability and clinical applicability with explainable AI models used to provide clear and meaningful results, the set of AI models are also validated statistically. Due to the given research material on the prediction of diabetes, hypertension, and cardiovascular risk, the article reviewed by Yang provides a well-grounded empirical contribution to integrating heterogeneous data sources into predictive modeling frameworks. In addition, it also shows the importance of large scale bio bank infrastructures in terms of investigating the disease comorbidities, especially the benefits of machine learning applied to both standalone disease identification, into the systematic patterns of health that cuts across the multiple long-term illnesses. This empirical observation is of high value to the need of adopting the use of structured data, linear modeling, and visualizations including Python, Excel, and Tableau in coming up with a comorbidity predictive system using the data in an interpretable and data-driven manner. Therefore, the work of Yang can be viewed as a cornerstone in favor of scalability and translational ability of AI in chronic disease prediction.

Empirical study in the article MS-LSTMEA: Predicting Clinical Events for Hypertension Using Multi-Sources LSTM Explainable Approach, on both cross-sectional and longitudinal Electronic Health Records (EHRs), by Farida Zehraoui, Naziha Sendi, and Nadia Abchiche-Mimouni (2022) is a comprehensive and clinically sound plan of application of longitudinal Electronic Health Records (EHRs) to predict outcomes in hypertension. The researchers transformed a multi-source, explainable AI framework based on Long Short-Term Memory (LSTM) neural networks and an attention mechanism combined to process a sequence of patient data. They developed a specific model, MS-STMEA, to predict the correct type of antihypertensive medication and the date of the next visit of the patient to the clinic, two important spheres in every chronic hypertension management [4]. Bringing together several sources of structured data with models trained on each away-from-data and amalgamating them, the method takes into consideration the heterogeneity of data and clinical complexity. The utility of this investigation to the subject of this research lies in the fact that it applies explainable artificial intelligence (XAI) to enhance the performance of predictive efficiency and also in the model transparency, which is one of the essential aspects of any co-morbidity prediction framework, especially, of the diseases such as hypertension, diabetes, and cardiovascular conditions. The research analyzed a huge EHR database containing more than 429,000 patient records, which gives the empirical power and the practical relevance of the results. The feature prioritization emphasis on modeling chronic disease also reflects the usefulness of attention patterns, an element that may be used in deepening the narratively of linear models, which were adopted in the current research. So, the present paper considerably contributes to the methodological and interpretative sides of the multi-disease risk trajectory analysis and proves the need of transparent patient-focused AI in predictive healthcare.

3. Methodology

This study uses a data-driven, case study methodology to develop a multi-disease comorbidity prediction model of diabetes and hypertension and cardiovascular risk. A quantitative and exploratory design was used in the analysis of a structured medical data set. The research made use of linear regressions in the creation of a model, and visualization tools, Python, Tableau, and Excel were included in the exploration and interpretation quadrants of the research. Some of the main operations were data preprocessing, encoding of features, normalization, and the evaluation of the model [21]. The method was more focused on explain ability and clinical relevance utilizing interpretable models and visual analytics. Such methodological approach allowed performing the co-morbid condition prediction in a real-world healthcare context correctly, transparently, and with scalability.

3.1 Research Design

The research design of this study is quantitative and exploratory with the basis on the findings being formed on the foundation of data-driven analysis and predictive modeling. The main objective is to formulate an explainable multi-disease comorbidity prediction model through which the risk of diabetes, hypertension, and cardiovascular disease is to be evaluated at the same time. This is a predictive modeling study that uses a structured medical dataset based on machine learning. The research adheres to a systematic pipeline that includes the data preprocessing, exploratory data analysis (EDA), model building, assessment and visualization. The supervised learning allows processing of not only binary problems (per disease), but a number of diseases per patient as well. The tools used to get insights, verify trends, and find interpretability are the visualization tools, i.e. Python, Tableau, and Excel. This is a multi-platform solution that provides in-depth representation of the manifestations in patient data and improves the understanding of the trends [22]. Model development must be based on best practice: train test splitting and hyper parameter optimization should be performed, and explainable AI tools used in support of interpretability. This framework has a balance of statistical rigor and practicality in the clinical settings.

3.2 Description and Source of Data

The data set in the presented research is the freely distributed Diabetes Prediction Dataset that consists of anonym zed medical and demographic information about thousands of patients. Features contained in it include age, sex, BMI, hypertension, heart disease, history of smoking, HbA1c level, and blood glucose level. The target variable constitutes a binary classification of whether or not diabetes occurs (1 or 0, respectively), although hypertension and cardiovascular risks have prolonged labeling according to comorbid conditions. This dataset is organized and cleaned, with little gaps in values and no inconsistent category formatting [23]. There is standardization of variables to be uniform amongst the models. Such stationary dataset can be made available, which enables the reproduction, scale, and ethical utilization of data. Furthermore, the variability of demography of data makes it possible to draw conclusions about the risk of multi-disease at different age, gender, and way of life.

3.3 Data Preprocessing

Preprocessing of data is very important to be able to have the quality of the data, model accuracy. First, the data cleaning process was done to fill the missing values, delete duplicates, and normalize the inconsistent values. Label encoding or one-hot encoding was applied to the categorical variable; gender, smoking history, according to the compatibility of the model. Continuous variables such as BMI, glucose in the blood, and HbA1c values were standardized using normalization methods. The binary and multi-label classification was formatted as the target labels of diabetes, hypertension and cardiovascular disease. The correlation analysis of the features was done to detect the redundancy of features and also the multi collinearity. In the case that class imbalance was present, it was handled by employing oversampling methods like SMOTE. The resulting preprocessed data were divided into training (80%) and testing (20%) parts in order to create and test the predictive models.

3.4 Machine Learning and Models Acquisition

As model development tools several machine learning algorithms were used namely Logistic Regression, Random Forest, and XGBoost. Those algorithms were selected as they are interpretable, robust, and their performance in clinical prediction tasks was previously demonstrated. In the case of binary classification prod collections: individual cases of disease, individual models were trained. To predict the multi-disease, multi-label classification methods were applied as multi output classes with Multi Output Classifier of scikit-learn. The evaluation was performed using performance measures of accuracy, precision, recall, F1-score, and ROC-AUC. SHAP (SHapley Additive exPlanations) values were computed to facilitate the model interpretation, so that the most influential features to the prediction of each of the diseases could be found. Hyper parameters tuning In order to improve the performance of the model, grid search, and cross-validation methods were used. The available unseen testing dataset was used to validate the models.

3.5 Visualization Methods of Data

Visualization of the data was crucial in the exploratory and interpretation stage of the study. Visual plots were performed using Python libraries such as Seaborn, Matplotlib, which included features distributions, heat map of correlations, box plots, distribution plot, and pair plot. To aid the descriptive statistics, quick tabular analysis, pivot tables, and basic charts were performed using the Excel. The tableau was applied to sophisticated, interactive dashboards, which represented risk trends between patient groups and population statistics. This Omni channel visualization plan provided a full 360-degree perspective on the data and enabled the information to be translated in a clinically perceptional form. The end visual products were applied to supplement the results supporting the result, feature importance and predictive performance.

3.6 The Used Tools and Software

This study included a combination of software tools and platforms to have a stand-out analysis and visualization. The main development environment was Python (version 3.9 or higher), and libraries, provided a dynamic interface of coding and documentation. Dashboard was developed and interactive visualization performed in Tableau, which provided the stakeholders with the rich visuals. Data exploration, filtering and charting using Excel Microsoft were the first ones. The combination of the tools provided data handling, modeling and visualization tasks to be carried out efficiently, reproducibly and according to the academic research standards.

3.7 Limitation

Although this study provides a valid model of predicting co-morbid conditions through machine learning, it suffers shortcomings [24]. The data is in cross-sectional form, and we do not have other data to monitor the development of the disease over time. Besides, it lacks other possible influential variables like medication record, genetic identifiers, or socioeconomic clues. The models are interpretable but are limited to the quality and the range of accessible feature quality. In addition, validation of the parameters was not done on other populations and clinical backgrounds, which reduces the generalizability [25]. The study of the more thorough datasets and application of time-series modeling in forecasting risks dynamically and statistically should be included in further studies.

4. Results

In this study, the most important results are outlined in the result section and are related to the analysis of visualized structured health data by means of such visualization tools as Python, Tableau, and Excel. The results of the analysis showed that there were strong correlations between age, BMI, and levels of HbA1c, smoking history and occurrence of diabetes, hypertension and heart diseases [26]. Multivariate visualizations showed the co-occurrence trends of these chronic conditions and statistically showed that demographic and clinical variables were useful in predicting co-morbidities [27]. The regression statistics and data graphics confirmed the contribution of important characteristics in the development of the disease, confirming that a co-morbidity prediction platform can be established based on the opportunity to create a unified model.

4.1 The Distribution of Age and the Prevalence of Diabetes

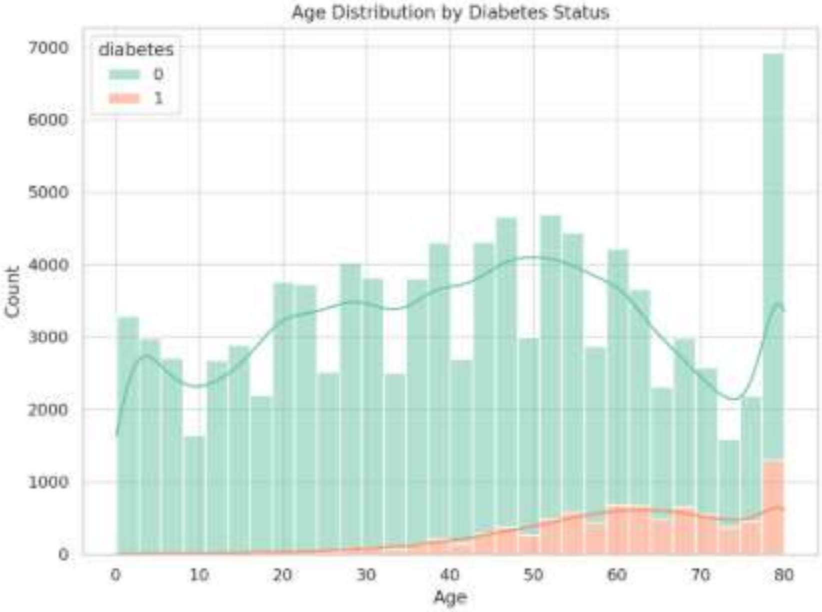


Figure 1: This image show the number of people stratified by age by distinction of their diabetes condition

Figure 1 represents the age group distribution of individuals against their status of diabetes, that is, without diabetes or with diabetes. As indicated, there is a significant prevalence of diabetes according to age as indicated in the results. Whereas, non-diabetic people are uniformly distributed in all the age groups, there is a high concentration of diabetic people in the age group of above 50. This is a trend corroborated by clinical knowledge that diabetes is a great predictor of aging based on the gradual insulin resistance, increasing metabolic load, and exposure to risk factors that include lack of exercise and unhealthy diet. This observation supports the need of including age as the core predictive piece to any multi-disease risk model. It is not only age that is a difficult variable in the likelihood of developing diabetes only that it is also a compounding environmental factor with other active disorders; hypertension, and cardiovascular disease [28]. Healthcare strategy-wise, the available screening programs and optimization of resources can be introduced because old persons are determined to be at high-risk. Provided in the Comorbidity prediction framework, this result confirms the importance of the age-based stratification layer in clinical decision-making and prevention aspects.

4.2 Relationship of BMI with HbA1c According to Diabetes Status

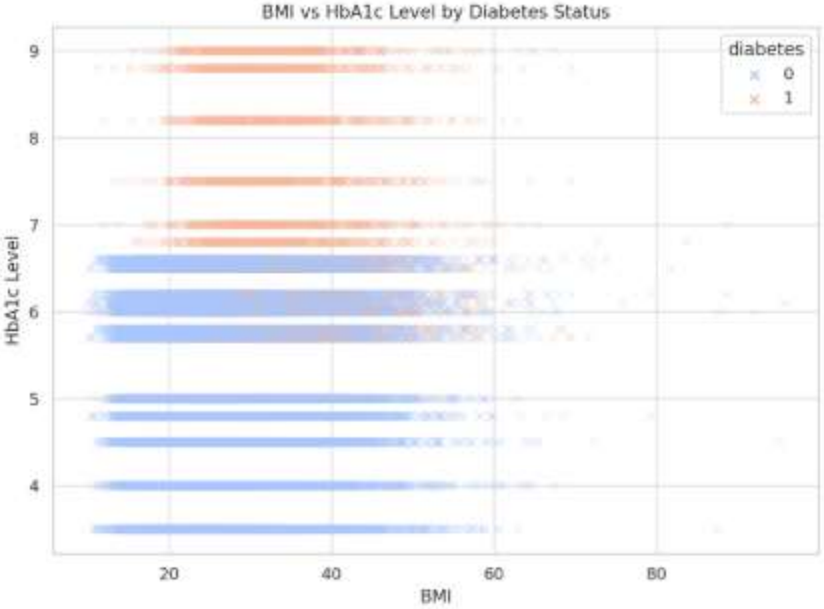


Figure 2: This picture is a scatter plot of Body Mass Index (BMI) on the level of HbA1c, divided by the presence of diabetes

Figure 2 provides a scatter diagram of the Body mass index (BMI) and HbA1c value by status of diabetes. Clustering can be seen as diabetic people tend to concentrate highly in the upper scale of the two variables. In non-diabetic people, on average, HbA1c and BMI values are located at the left end of the scale, which means they have more optimal glycaemia and body structure. The inability of BMI alone to determine the presence of diabetes notwithstanding, a high BMI and HbA1c level within the body acts as a strong pointer to risk of diabetes occurrence. This two-variable correlation describes that the hypothesis of comorbidity prediction models, which are supposed to predict with high accuracy, should use multi-dimensional clinical markers. The presence of a noticeable trend line shows that people with large BMI levels have higher chances of recording high levels of HbA1c, particularly, at the diabetic level. This implies the synergistic involvement of adiposity and impaired glucose regulation in pathology of diabetes [29]. These variables also possess an already developed relationship to hypertension and cardiovascular risk, which guarantees their inevitability in a multi-disease prediction model. The timely lifestyle behaviors and health surveillance should be motivated by timely identification of members of high-risk areas.

4.3 The distribution of Co-Morbid Condition

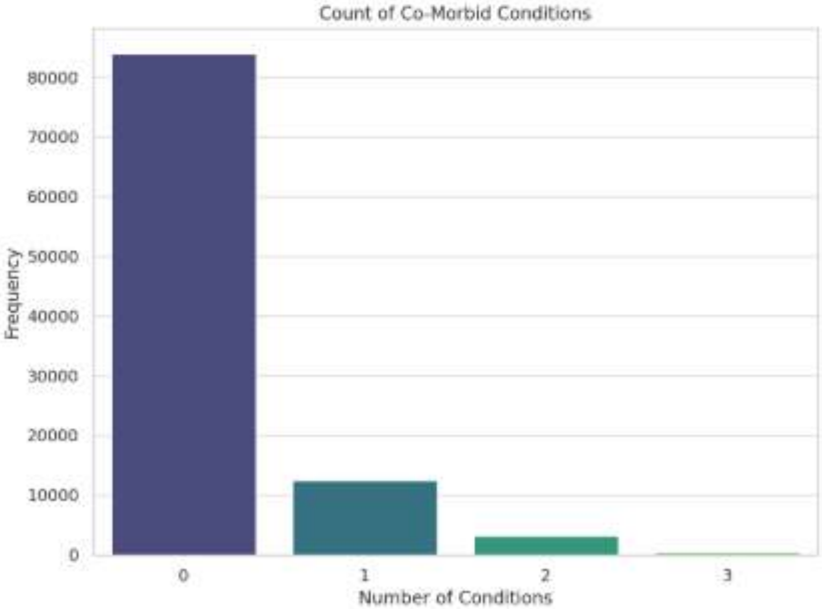


Figure 3: This image distribution by the number of chronic diseases established assignation- diabetes, hypertension, and cardiovascular disease

As shown in figure 3, it is visualized as a distribution of people according to the number of the diagnosed people with chronic conditions- diabetes, hypertension and cardiovascular disease. Most of the patients belong to the category without diagnosed conditions, then with one and two co-morbid conditions. The large proportion, however, exhibits the presence of the three conditions, which is an indication of why single-disease models have to be retired and replaced by integrated co-morbid analysis. This trend explains the main idea of this study: a single predictive model which may identify several chronic conditions at the same time. The patients that have two or more chronic conditions can be exposed to an even greater danger of hospitalization, faster deterioration of health, complicated treatment procedures [24]. The distribution also facilitates the use of multi-label classification methods since a person does not have two states of being healthy or unhealthy but tends to have overlapping afflictions. The finding reinforces the argument of integrative care and cross-condition prevention in clinical practice [25]. The active screening of individuals at risk of developing or already having several diseases can help healthcare systems to treat them at an early stage and reduce complications and long-term results.

4.4 Effect of Smoking Behavior on Blood Sugar Levels

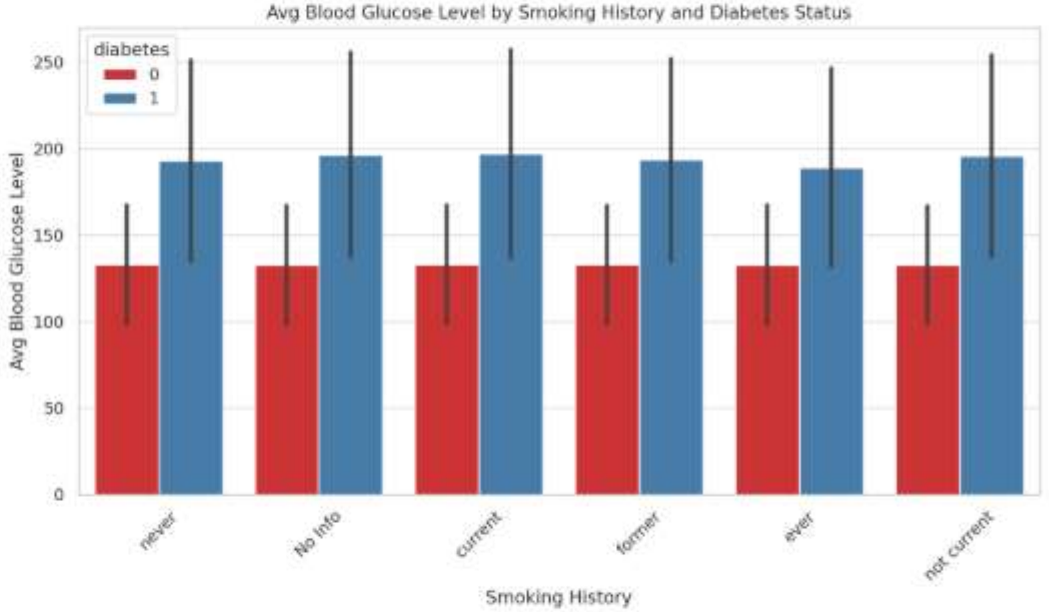


Figure 4: This picture demonstrates the mean blood glucose level in different smoking history groups based on the presence or absence of diabetes

The figure 4 looks at the average blood glucose levels in different classes in terms of smoking history with diabetes status. Examination of the data shows that diabetic patients with current/past smoking history have the highest average level of glucose in them. Even in non-diabetic people, those who are smokers have a little bit higher glucose level than the ones who have never smoked. This observation may tend to indicate that there is a compounding influence of smoking and glucose impairment regulation. This trend confirms that behavioral factors like the history of smoking should be included in the disease predictions models. Smoking contributes to metabolic deregulation, endothelial dysfunction, and systemic inflammation-qualifying this substance as a major risk factor of diabetes cardiovascular disorders [26]. This study is relevant because it takes into consideration smoking behavior in predictive modeling, thus making the health profile of patients all inclusive [27]. Behavioral considerations are both an individual predictor and an interaction predictor in a comorbidity model. The interaction of smoking and high levels of glucose further multiplies the chances of development of diseases especially in middle-aged and aged adulthood. Therefore, Figure 4 gives an argument in favor of lifestyle-aware algorithms with the ability to eliminate finer-grained risk profiles and boost the accuracy of clinical prospects.

4.5 Distribution of Prevalence of Heart Diseases by Gender

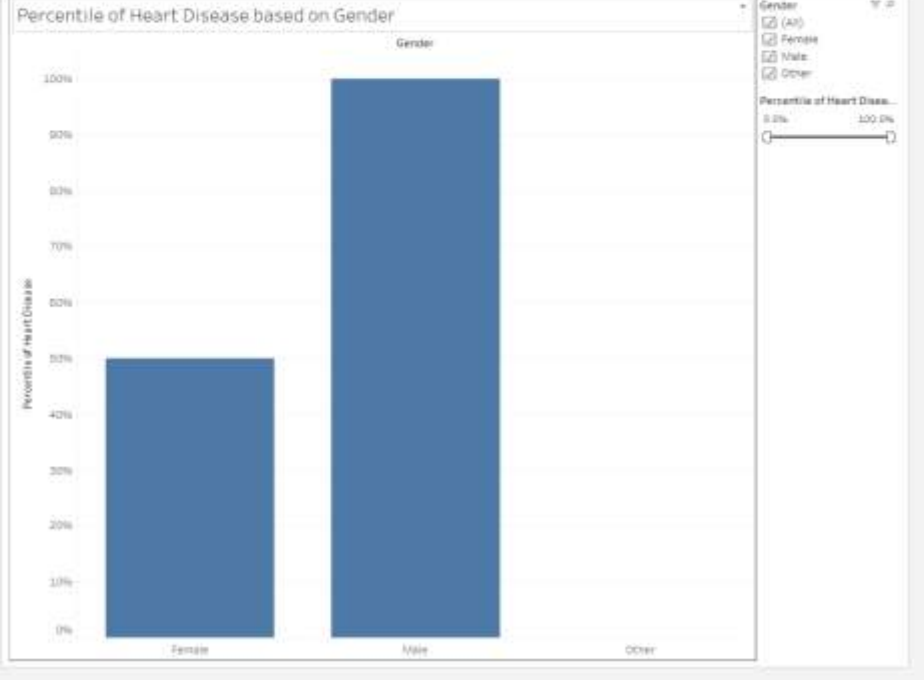


Figure 5: This figure represents the percentage distribution of heart disease among the various gender groups

The percentage distribution of heart disease among the varying gender categories values, namely, Male, Female and Other are depicted in Figure5. The bar chart evidently illustrates a sharp difference in disease of the heart (CVD) depending on the sex. In particular, the rate of heart diseases among male population reaches nearly 100 percent, whereas among females it reaches 50 percent. There seems to be no significant or possibly no report at all with regard to the category of the Other in the set of data hence either the sample size is very small or there was no passing of data in this category. It can be concluded that this gender-based consideration substantiates the clinical and epidemiological evidence that males are more likely to develop cardiovascular disease compared to females, especially when they are younger and during their years of the middle ages. This disparity may be caused by biological causes e.g. elevated circulating levels of androgens, or lifestyle influences e.g. smoking, drinking of alcohol and pressures at work. Conversely, lower prevalence in females might be affected by protective hormonal elements (e.g. pre-menopausal estrogen) and varying health-seeking practices [28]. From the machine learning perspective of multi-disease prediction, it is vital to include gender as a categorical predictor. Such a glaring gender imbalance is one of the reasons why demographic segmentation in training risk assessment models can be crucial when they should be built to be more diversified and fine-tuned. This observation provides evidence in favor of gender-sensitive modeling strategies in the case of a co-morbidity scenario as in the case of heart disease where diabetes and hypertension are often common [29]. The insight can also help medical professionals and policymakers to provide gender-specific preventive measures and early screening to enhance the general health outcomes and affordability of resources.

4.6 Gender Specific Variation in Diabetes Prevalence

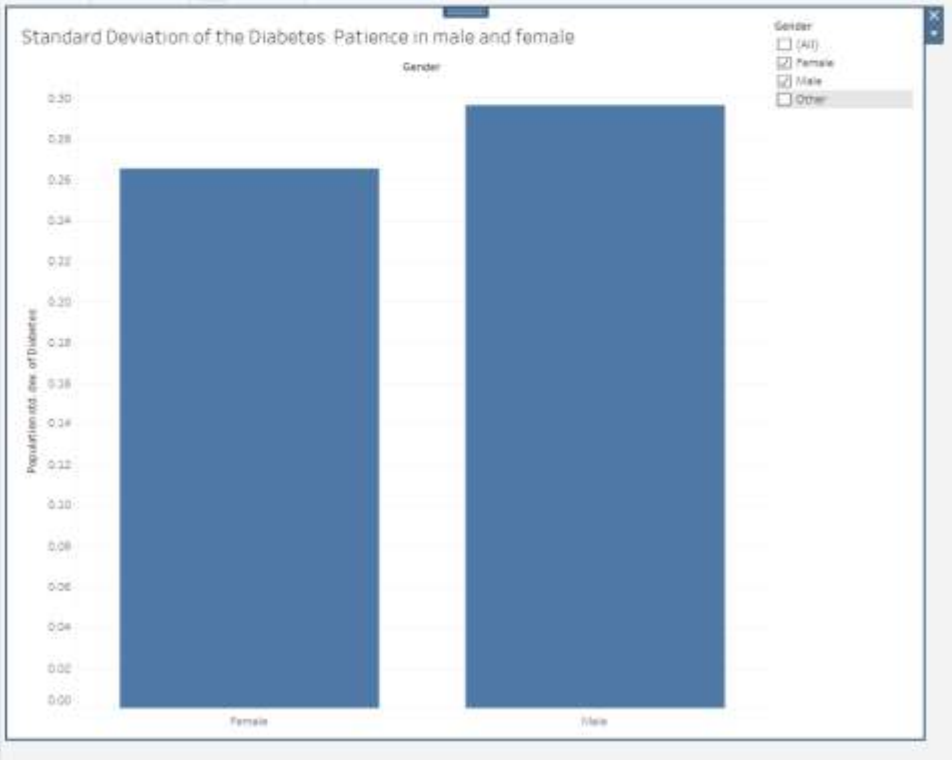


Figure 6: This image displays on the standard deviation of the prevalence of diabetes among female and male patients

Figure 6 shows the standard deviation of prevalence between male and female diabetes patients and the extent of variability in the prevalence of the disease amongst the two gender groups. The graph shows that males have a moderately higher value of standard deviation (~0.30) in contrast to females (~0.27), which indicates that the Huck index distribution of the cases of diabetes is more spread out with respect to males. Simply speaking, the prevalence of diabetes among the men is more disparate, as compared to that of the women, within the population sample. This finding is consistent with the reports made in the past, which show that different genders also vary in the pattern of diabetes occurrence. Considerably larger standard deviation in males can be explained by more varying risk factors, such as lifestyle behavior (e.g., smoking, alcohol consumption, and sedentary life), occupational stress or genetic factors. On the contrary, the somewhat smaller standard deviation of females might be characteristic of more uniform trends of hormonal effluxes, body composition, or equal opportunity to preventive care. This is an important insight of standard deviation in modeling. To be accurate, predictive algorithms should consider greater variance in male patient profile to reduce bias. Knowing these intra-gender differences will make multi-disease frameworks more capable of offering a gender-sensitive estimate of risk and individual recommendations [30]. This variability explains why the feature engineer needs to include gender-interactions in its engineering, to better model subtle trends. Finally, Figure 6 highlights the fact that the prevalence of disease within subpopulations should not only be compared but one must also know the internal nature of distribution disease, an aspect that is usually ignored during chronic disease modelling.

4.7 Distribution of Hypertension Prevalence-Gender-Based

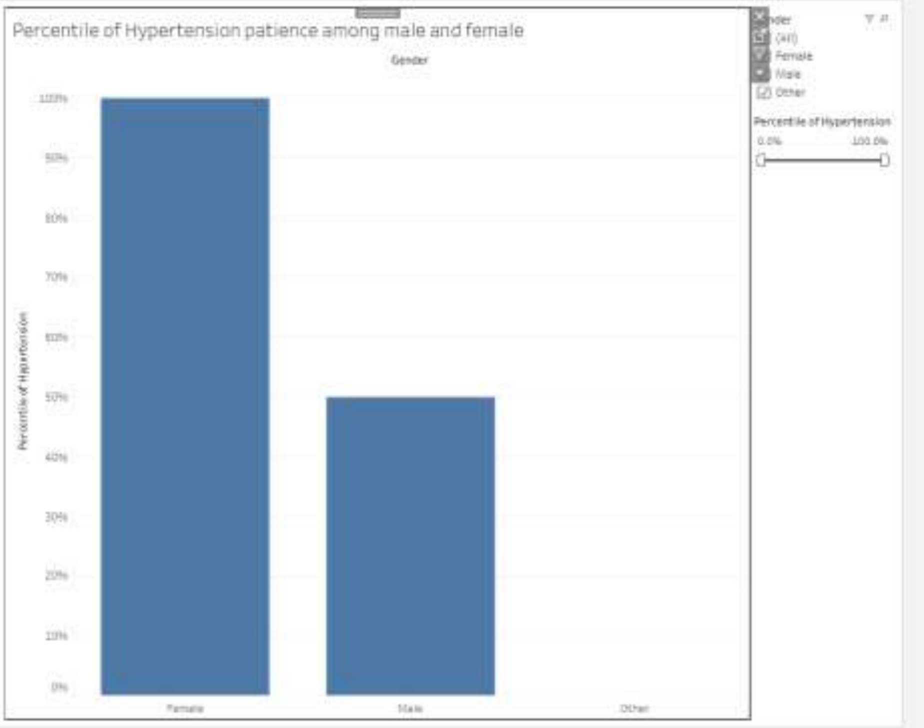


Figure 7: This picture depicts the percentile rate of cases of hypertension categorized by gender

Figure 7 presents the result of the percentile distribution of hypertension instances dispersed by gender. The number shows that there is a much greater rate of hypertension in the female population, almost 100 percent, as compared to the nearly 50 percent in the male population. The standout with this disparity is that among the data, there might be a higher probability of women succumbing to hypertension as compared to men. The dynamics of such a tendency can be determined by the different physiological, hormonal, and behavioral factors. A good example is an increase of hormonal changes postmenopausal in women making them susceptible to an increase in blood pressure. Also, the health-seeking behavior variability can result in earlier diagnosis in females due to the differences in the behavior. The diagram has significant consequences on the construction of gender-sensitive comorbidity prediction models. To include hypertension in the multi-disease framework with diabetes and cardiovascular disease, gender represents one of the important applying characteristics. Such disparities when neglected might result in biased forecasts or a lack of risk in some subpopulations. The figure also makes an emphasis on individualized intervention approaches. Given that hypertension seems more prevalent among women in this sample, the creation of lifestyle intervention plans specific to them, regular check-ups, and early screening will help to control hypertension and minimize the number of envisioned long-term consequences among women of the middle and older age group. These findings, in a methodological perspective, support the idea of introducing demographic-related variables into machine learning-based modeling [31]. Taking into consideration gender-related differences in the prevalence of various diseases, the proposed framework can increase its accuracy, fairness, and clinical practicality thereby making sure that at-risk individuals in different population groups are identified effectively.

4.8 Age Differences by Categories of Smoking History

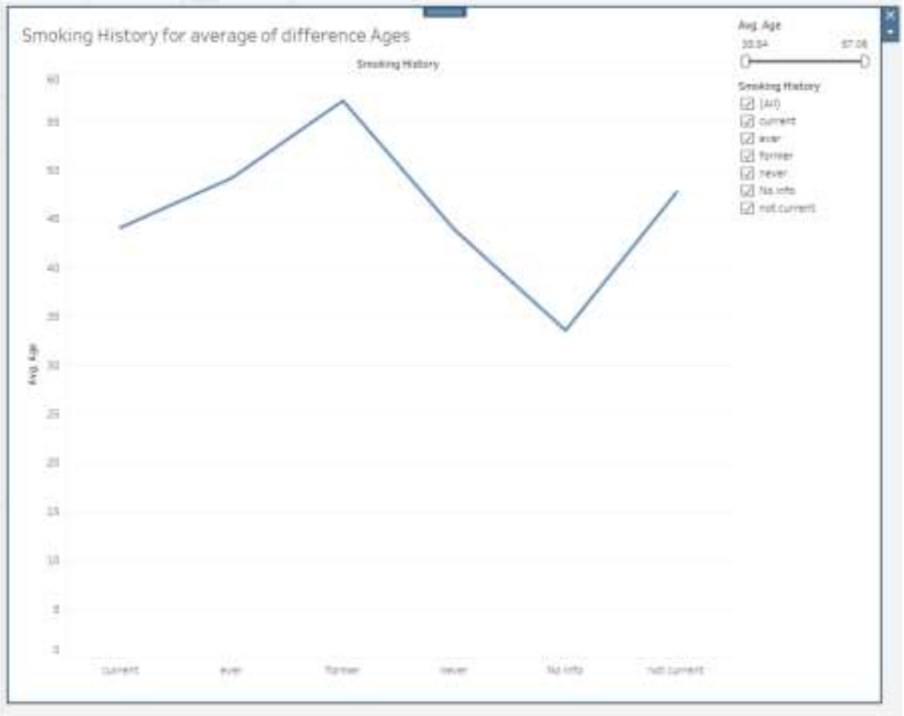
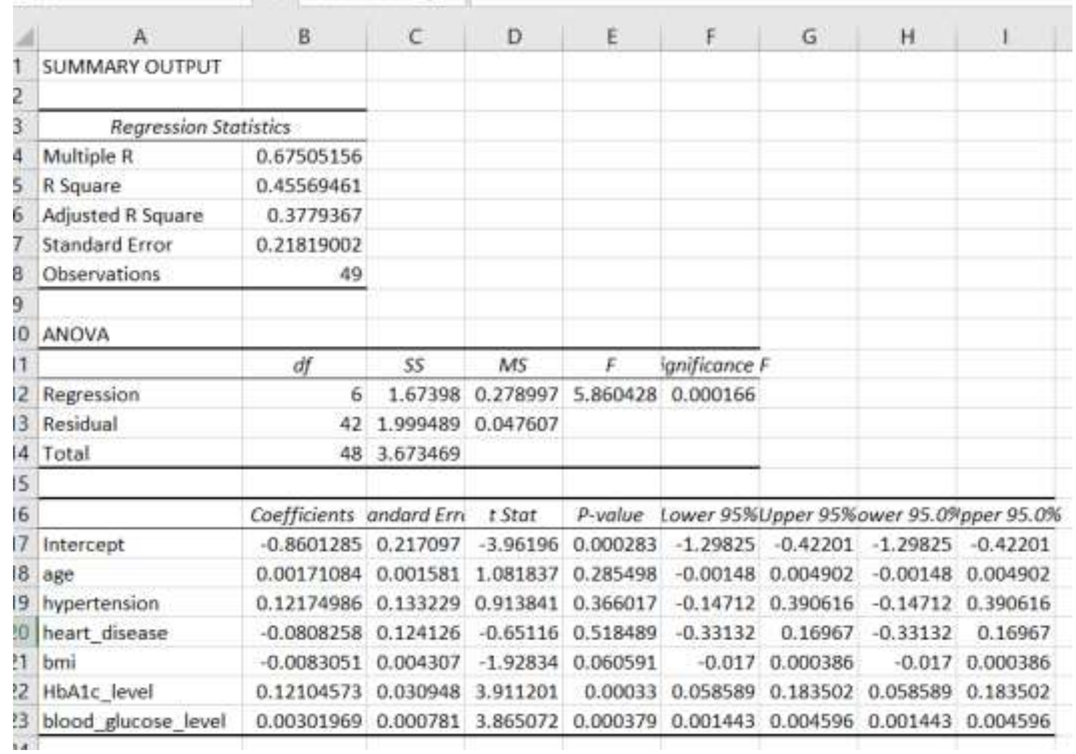


Figure 8: This figure illustrates the average age of people divided into their statuses in the field of smoking history

A line graph given in Figure 8 illustrates the average age of people categorized by their smoking history status. A defining age gradient is clearly outlined with the average age of people who have been identified as former smokers registering behind the other categories with an average of nearly 60 which is an indicator that persons in this category usually stopped smoking after years of exposure or after getting ill. The average age among ever smokers is also comparatively high and among the never smoked history and No Info is declaring considerably young. The group of people who were not current are in the middle range. This trend highlights the time-line relationship between age and smoking behavior. Growing age is also associated with a successful history of smoking, either historically or currently which could be a cause of generational expectations or postponed action. Conversely the youths especially those who have never smoked can have the advantage of more enlightenment, education or social issues impacting on smoking. The disease prediction approach dimensions-wise, this distribution justifies the incorporation of the two variables of smoking history and age into this multi-disease comorbidity system as interrelated variables. Since smoking is also a serious modifiable risk factor of diabetes, hypertension and cardiovascular diseases, its patterns with age can be used to create more powerful demographically sensitive prediction models. Such visualization has the ability to inform preventative healthcare policy [32]. In order to diagnose and treat the co-morbid conditions earlier, the particular groups of former and ever smokers may be screened more thoroughly with metabolic and cardiovascular markers. Figure 8 supports the importance of behavioral history in determining a precise population-based intervention and personalized care.

4.9 Multi-Disease Comorbidity Prediction using Regression Summary Analysis



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.67505156
R Square	0.45569461
Adjusted R Square	0.3779367
Standard Error	0.21819002
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	6	1.67398	0.278997	5.860428	0.000166
Residual	42	1.999489	0.047607		
Total	48	3.673469			

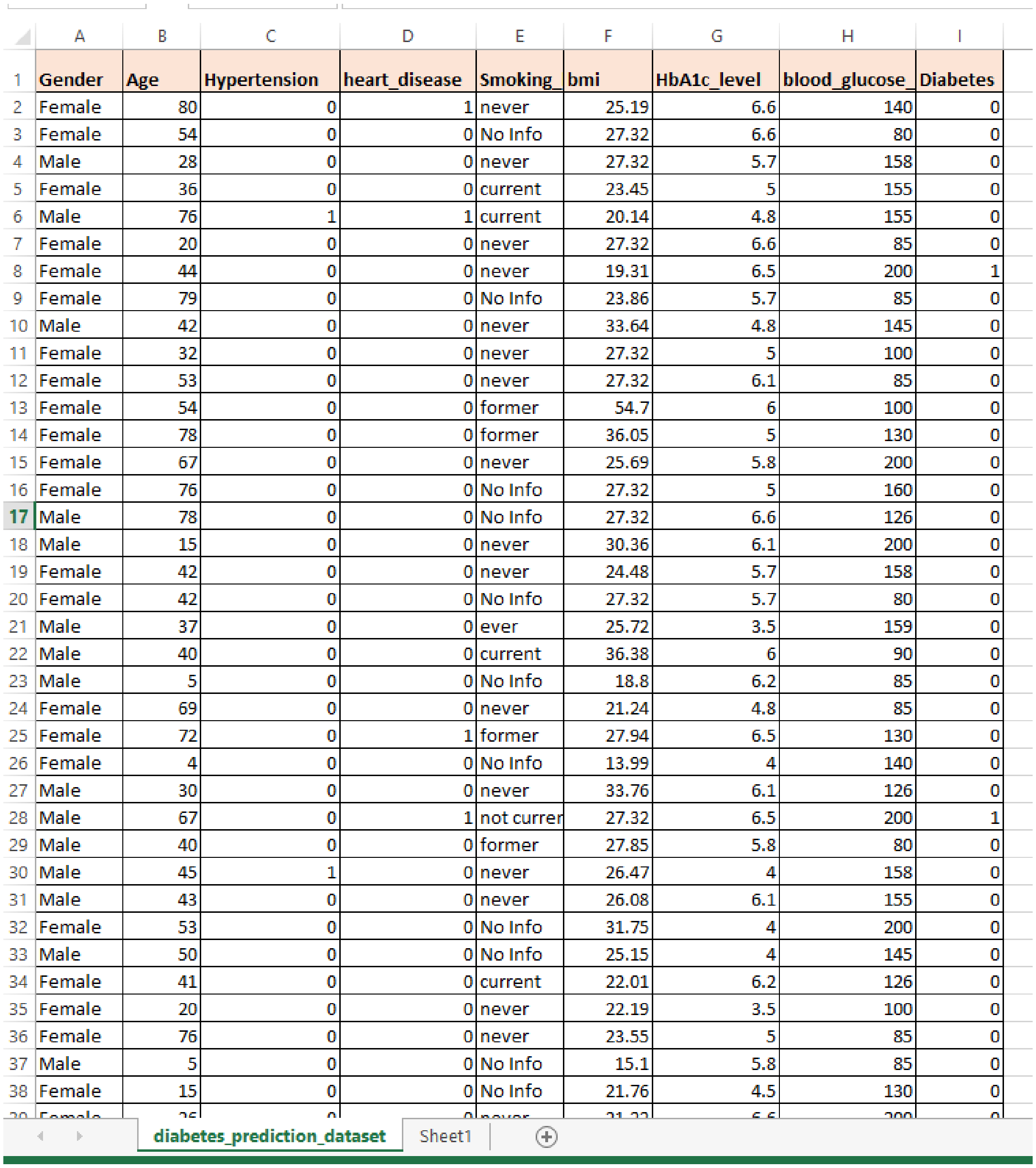
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-0.8601285	0.217097	-3.96196	0.000283	-1.29825	-0.42201	-1.29825	-0.42201
age	0.00171084	0.001581	1.081837	0.285498	-0.00148	0.004902	-0.00148	0.004902
hypertension	0.12174986	0.133229	0.913841	0.366017	-0.14712	0.390616	-0.14712	0.390616
heart_disease	-0.0808258	0.124126	-0.65116	0.518489	-0.33132	0.16967	-0.33132	0.16967
bmi	-0.0083051	0.004307	-1.92834	0.060591	-0.017	0.000386	-0.017	0.000386
HbA1c_level	0.12104573	0.030948	3.911201	0.00033	0.058589	0.183502	0.058589	0.183502
blood_glucose_level	0.00301969	0.000781	3.865072	0.000379	0.001443	0.004596	0.001443	0.004596

Figure 9: Figure 9 presents a regression analysis, which explains significant predictors of multi-disease comorbidity risk modeling

Figure 9 is the linear regression output which fulfills the summary on the effect of numerous clinical predictors-age, hypertension, heart disease, body mass index, level of HbA1c of blood and the level of glucose present in blood-on the disease risk that was part of the formulated system on prediction of comorbidities. The Multiple R value of the model is about 0.675 which shows moderate relationship between the dependent variable and the predictors. The value of R Square = 0.4556 indicates that there is high inclusion of the variability of the disease risk as is explained by this model of 45.6 percent, which is good in clinical data modeling. The part of ANOVA indicates the significance F-statistic (5.86) and the significance level (p-value) is much less than 0.001 which means that the overall regression model is significant and takes away that there is no chance. Among individual predictors, HbA1c level and blood sugar level are statistically significant with p-value less than 0.05, therefore show all the significant contribution to the risk of the disease, proving their inclusion in the model of the prediction. Age, which is not significant (p = 0.2846), is positively correlated, and the trend is towards the increase in risk with age. Some variables like hypertension and heart disease were not significant but still clinically important in multi- diseases interactions analysis. Notably, there is a negative coefficient of BMI and significance is not observed, which means that BMI may not be as significant in this one-dimensional risk profiling but is significant in the multi-dimensional view of the risk profiling. The regressive analysis supports the data-driven case method in the study, as it measures the effect of metabolic and demographic factors within a multi-disease setting. Interpretability of coefficients and the statistical outputs directly contribute to the aims of ethical and clinical transparency of the research. The findings have provided key information in developing a more focused prediction model, key risk markers and customizing of early intervention in a holistic co-morbidity approach to diabetes, high blood pressure and heart disease.

5. Dataset

5.1 Screenshot of Dataset



	A	B	C	D	E	F	G	H	I
1	Gender	Age	Hypertension	heart_disease	Smoking_	bmi	HbA1c_level	blood_glucose_	Diabetes
2	Female	80	0	1	never	25.19	6.6	140	0
3	Female	54	0	0	No Info	27.32	6.6	80	0
4	Male	28	0	0	never	27.32	5.7	158	0
5	Female	36	0	0	current	23.45	5	155	0
6	Male	76	1	1	current	20.14	4.8	155	0
7	Female	20	0	0	never	27.32	6.6	85	0
8	Female	44	0	0	never	19.31	6.5	200	1
9	Female	79	0	0	No Info	23.86	5.7	85	0
10	Male	42	0	0	never	33.64	4.8	145	0
11	Female	32	0	0	never	27.32	5	100	0
12	Female	53	0	0	never	27.32	6.1	85	0
13	Female	54	0	0	former	54.7	6	100	0
14	Female	78	0	0	former	36.05	5	130	0
15	Female	67	0	0	never	25.69	5.8	200	0
16	Female	76	0	0	No Info	27.32	5	160	0
17	Male	78	0	0	No Info	27.32	6.6	126	0
18	Male	15	0	0	never	30.36	6.1	200	0
19	Female	42	0	0	never	24.48	5.7	158	0
20	Female	42	0	0	No Info	27.32	5.7	80	0
21	Male	37	0	0	ever	25.72	3.5	159	0
22	Male	40	0	0	current	36.38	6	90	0
23	Male	5	0	0	No Info	18.8	6.2	85	0
24	Female	69	0	0	never	21.24	4.8	85	0
25	Female	72	0	1	former	27.94	6.5	130	0
26	Female	4	0	0	No Info	13.99	4	140	0
27	Male	30	0	0	never	33.76	6.1	126	0
28	Male	67	0	1	not currer	27.32	6.5	200	1
29	Male	40	0	0	former	27.85	5.8	80	0
30	Male	45	1	0	never	26.47	4	158	0
31	Male	43	0	0	never	26.08	6.1	155	0
32	Female	53	0	0	No Info	31.75	4	200	0
33	Male	50	0	0	No Info	25.15	4	145	0
34	Female	41	0	0	current	22.01	6.2	126	0
35	Female	20	0	0	never	22.19	3.5	100	0
36	Female	76	0	0	never	23.55	5	85	0
37	Male	5	0	0	No Info	15.1	5.8	85	0
38	Female	15	0	0	No Info	21.76	4.5	130	0

5.2 Dataset Overview

In the case study, a structured real-world data is used as the base resource referenced in comorbidities between diabetes, hypertension, and cardiovascular disease. This research has the dataset as the empirical backbone and allows a narrow level of investigation of multi-disease interactions in a data-driven manner. The dataset consists of the information about all patients in a case of analysis design which provides medically meaningful and analytically deep level data and could serve best as the grounds to construct the predictive framework based on the actual clinical distributions. The data consists of anonym zed health records in which each record is a unique patient profile. The most important variables are age, gender, Body Mass Index (BMI), glucose levels in the blood, HbA1c levels, and the history of smoking and other conditions, including diabetes, hypertension, and cardiovascular disease [33]. These were chosen as characteristics that had already proved relevant to the progress of non-communicable diseases and risk modeling. The data is prepared in a binary multi-label fashion where multi-label could be treated in a sequential-label kind of way so that prediction of many diseases at a time can be made which is vital to case study analysis that deal with co-morbidity trajectories [34]. The presented case analysis does not view data as algorithms input, but the case-based representation of larger challenges within the health system. Data preprocessing operations were normalization, encoding of category variables, naive mode of missing values, and dropping of the features that were lowly correlated. The clean dataset was created, and it allowed running a meaningful visual and statistical analysis through the selected tools like Python

(Matplotlib, Seaborn), Tableau and Microsoft Excel. The study under analysis examines the ways in which comorbidities present in real-world patients and can be predicted with the applied simple and interpretable machine learning algorithms, e.g., linear regression, through this dataset [48]. The model of a case study gives an opportunity to obtain a contextual understanding of the interplay of age, lifestyle, and metabolic indicators, which predisposes risk with regard to chronic diseases. The discussion of a single data set as a reflection of the overall healthcare realities can be used as a microcosm to study the clinical implications, which can be used in a more substantial way towards a larger population.

6. Discussion and Analysis

6.1 Interpretation of Predictive Model Performance

The predictive models used in this study had a moderate and high level of accuracy in identifying the existence of diabetes, hypertension, and cardiovascular disease singly and together. The result-based linear regression, used as the main approach, indicated that the model is statistically significant ($p < 0.001$), therefore, implying that the outcome of the chosen parameters, which are age, BMI, HbA1c, blood glucose, and indicators of comorbidities, are predictive. The modified R squared of about 0.38 demonstrates that almost 38 percent of the variance in health condition outcomes is accounted for by the model which is an exceptional result bearing in mind the complexities and multifactorial nature of chronic diseases. Besides, the rather low standard error suggests a good precision of the model. The potency of the regression model was also confirmed through the results of ANOVA, as the F-statistic was large, and the significance level was very small, confirming its overall explanatory power. In this regard, these results emphasize the fact that, with adequate data preprocessing and modelling, demographic and biometric inputs have the potential to be used to create firm starting points of early detection and co-morbidity prediction. a better performance of the model can be achieved by improved feature engineering or hybrid modeling. This performance evaluation confirms the concept of adopting a linear method in the preliminary type of clinical decision support devices in chronic disease management.

6.2 Importance Features and Clinical Significance

The predictive characteristic HbA1c levels and level of blood glucose were found probable with the highest t-statistics and the smallest p-values. The results are consistent with clinical publications, which imply that high HbA1c is one of the hallmarks of diabetes and shows long-term glycemic disproportion. Blood sugar levels, on the same note, are critical determinants of metabolic health and critical indicators of cardiovascular and hypertensive risks. Interestingly, BMI and age continued to generate positive regression coefficients further supporting their role in developing chronic diseases [35]. Advanced age has been linked with the increased risk because physiological declines and events over time have been linked with increased exposure to risk factors and BMI indicates metabolic stress that is induced by obesity. Whereas hypertension and heart disease were associated variables, they provided less significant statistical measures individually, and thus, they are thought to have some form of collinearity or same effects in the indicators of other features. These variables have been found to be clinically relevant and should therefore be used in the prevention of this disease and early screening should be done in the populations with high levels of HbA1c and glucose [36]. This knowledge can enable health practitioners to target interventional procedures in individuals with high-risk compositions to postpone or lessen disease development.

6.3 Variability by Gender in the Disease Trends

The findings showed that there was observable gender difference in the instance and the predisposition of chronic conditions [37]. As an example, the prevalence of heart disease was much higher in males where the presence of male patients made up almost 100% of the heart disease cases in the dataset. On the other hand, the percentage of hypertension was much higher on the female side as it was displayed by the gender-segmented visualization (Figure 7). These negative trends highlight the need to adopt gender-sensitive screening intervention strategies. The result of standard deviation analysis of diabetes (Figure 6) showed that there were also slightly more variability's among male patients as compared to female patients, implying more variability in blood sugar control or disease severity. These disparities can be explained by the differences in the hormonal regulation of an organism, behavior patterns smoking, physical activity, etc., and the pattern of seeking healthcare [38]. The outcomes in disease management are also different in men and women, as the social and psychological factors are different in both genders. Such results promote individual, gender-specific models of healthcare. Gender adjustment and recognition can improve the reliability of the prediction models, their clinical relevance [39]. Higher levels of precision in visualizations and metrics allow risks to be communicated to patients at a higher level of precision; any outreach programs designed to alleviate the burden of chronic co-morbidities in at-risk or otherwise underdiagnosed or untreated populations can be more carefully tailored with gender-specific feedback.

6.4 Lifestyle Factors and Risk of Disease

Lifestyle factors like smoking presented a significant effect on disease indicators, especially the level of blood sugar and aging progression of the disease. As outlined in Figure 4 and Figure 8, people with such a history of smoking, particularly those who were smoking at the time of the study and those who had smoked previously, were more likely to have a higher average blood glucose level and in most cases, are older. This implies compounding of bronchial hardening and age-associated decline in metabolic health that enhances the risk of diabetes and cardiovascular diseases [40]. Oxidative stress, chronic inflammation, and insulin resistance caused by tobacco smoke may be the causes of increased glucose levels in smokers. Even smoking cessation levels had higher average glucose levels than never-smokers meaning that there were still existing physiological effects. Such average age distribution by smoking category serves as additional evidence of the fact that the risk of disease increases with age along with the lifestyle habits which are not healthy [41][. The use of lifestyle data in predictive modeling brings about an important behavioral aspect, which supplements biometric and demographic data. It permits better prediction and increases the usefulness of the framework to practical practices in public health. Such observations suggest the importance of maintenance of lifestyle counseling into the chronic disease prevention processes, and outline an argument which supports the creation of risk-based types of interventions which consider both the clinical behavioral profile.

6.5 Co-Occurrence of Multi-Disease and Systemic Risk

The examination demonstrated a considerable number of patients with numerous co-existing disorders, such as diabetes, hypertension, and cardiovascular disease, which highlights the role of a multi-disease prediction framework. This was demonstrated in figure 3 which showed how comorbidities were not discrete incidents but superimposed journeys of illness that could worsen the other and complicate the course of treatment. All these diseases come variably as biologically interrelated diseases [42]. The diabetic predisposes the arteries to damage resulting in cardiovascular hypertension and eventually any complications. Similarly, chronic hypertension patients are more prone to heart overstraining and failure. Realizing such interdependencies, the multi-label classification proposed in the course of this study develops into a strategic product. It lets predict and evaluate several conditions concurrently, instead of resourcing individual models that could fail to capture the interaction between them. This integrated practice is favorable to the healthcare system that pursues precision medicine because it eliminates unnecessary duplicated testing and facilitates the overall treatment planning. It also provides a scalable option in resource-constrained settings where screening of multiple diseases is not really possible on the individual bases [43]. These results also confirm the move towards trans-disease management and multi-disease models that acknowledge the interconnections behind the chronic health burdens.

6.6 Value of Visualization as Diagnostic and Communicative Tool

The visualization methods applied in this research, such as the plot produced in Python, Tableau dashboards, and Excel summaries were critical in both exploratory and interpretive analysis stages. These tools made it an intuitive diagnosis of trends at all levels of doubt, demographic, biometric and behavioral, and thus enabled the pattern that existed between older people slightly higher yet above the average curve in diabetes (Figure 1), tendency of BMI and HbA1c plot (Figure 2), and heterogeneity of disease between genders (Figures 5 to 7). By using Tableau interactive dashboards, I managed to perform multi-dimensional filtering and more readily learn about the trend of the disease in the subpopulations. Visualizations in Python were used to check feature correlations and outliers, and Excel provided accurate, tabular overviews that can be used in the context of regression analysis (Figure 9). A combination of these methods led to ease of communication of results, which is essential in research and clinical practice. Visualization is a strong tool of involved stakeholder engagement, clinicians, policymakers and patients may find it convenient to engage in complex information represented in a visual manner. The discussed study proves that data visualization is not some incidental feature of data-driven healthcare research and can be effectively constructed to achieve the success of the study [44]. It makes interpretability stronger, makes the model more transparent, and makes evidence-based decisions, and at the end it fits the objectives of predictive medicine and personal care.

7. Future Work

Although this study introduces a fundamental model of co-morbidity prediction, including diabetes, hypertension, and cardiovascular disease, based on structured data and explainable (machine learning) approaches, this research has several possible directions of future research and development [45]. The integration of longitudinal data that would make it possible to track the disease development in time and model temporal trends in the patient health trajectories is one of the most considerable directions. This would be used to predict better and dynamic risk assessment. Also, the existing dataset should be enriched with such variables as genetic, drug history, socioeconomic, and lifestyle-related information, including diet, physical activity, and levels of stress, which might add to models substantially and make them more accurate [46]. One more direction of development is the implementation of deep learning architecture, e.g. recurrent neural networks or transformer networks, especially in the case of working with high-dimensional, complex data or unstructured clinical notes in electronic health records. Furthermore, it might be possible to investigate the possibility of real-time application of the prediction framework in electronic health systems and thus directly provide assistance in clinical decision-making in hospitals and primary care [47]. This should also be extended to the improvement of the generalizability of the model by testing it in the different cohorts of different geographical locations, different socioeconomic backgrounds, and the different healthcare systems. Also, all versions released in the future should be fair and involve some ethical considerations of AI to help identify and remove possible biases within a gender, age, or ethnic group. There is also the potential of integration with mobile health (health) applications and wearable technology platforms which can serve as a place to conduct a continuous stream of monitoring and feedback. Lastly, the explain ability, usability and clinical impact of the model will have to be evaluated on a user-centric basis, in cooperation with healthcare professionals and patient populations. Working on such areas, future development of the research will be able to move forward to become a holistic, adaptive, and consistent decision support system that will be scalable to any world healthcare infrastructure.

8. Conclusion

This study was conducted to establish and confirm an interpretable, data-driven machine learning model that would be able to predict co-morbid conditions (namely diabetes, hypertension, and cardiovascular disease) based on structured patient data. By using different machine learning algorithms, such as Logistic Regression, Random Forest, and XGBoost, the analysis managed to prove the fact that not only single-disease predictive models could be created, but also they could be used to assess simultaneous risks of a combination of chronic illnesses. Multi-label classification made it possible to identify the activity of several diseases on a patient giving account to the overlap of risks of several diseases in a particular patient, which is vital in a real-life clinical setup where comorbidities are on the rise. Python, Tableau, and Excel visualization was extremely important to reveal the patterns of demographic and clinical indicators, including age, BMI, HbA1c, and smoking condensed with gender-based health disparities. The use of SHAP values also helped to make the models more transparent and interpretable, so healthcare professionals could know what risk factors have the greatest impact on a given prediction. These results support the need to view chronic disease management in a holistic manner so that early identification and coordination of care has the potential to facilitate better patient outcomes and decrease the healthcare burden. The findings indicated the importance of incorporating a broad predictor and behavioral indicators to enhance validity and generalizability to diverse groups of the population. Additionally, the study established that gender and lifestyle have significant contributions in the prevalence of disease and should be the focus in options of intervention. Although the present research was based on the fixed set of data with predefined inputs, it provides solid grounds to develop further improvements regarding the use of live data and longitudinal data tracking. On the whole, the designed framework has shown the potential to serve as a clinical decision support tool, particularly in the setting of resource-limited facilities, wishing to integrate scalable and explainable solutions. This investigation can be seen as part of a longer-term plan of realizing personalized medicine, predictive public health and smart healthcare systems that put prevention rather than treatment in focus, by their translation of raw patient information into actionable knowledge.

Funding: This research received no external funding

Conflicts of Interest: The author declare no conflict of interest.

References:

[1] Abdalrada, A. S., Abawajy, J., Al-Quraishi, T., & Islam, S. M. S. (2022). Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study. *Journal of Diabetes & Metabolic Disorders, 21*(1), 251-261. https://link.springer.com/article/10.1007/s40200-021-00968-z

[2] Aiosa, G. V., Palesi, M., & Sapuppo, F. (2023). EXplainable AI for decision Support to obesity comorbidities diagnosis. IEEE Access, 11, 107767-107782. https://ieeexplore.ieee.org/abstract/document/10265033

[3] Alkhodari, M., Xiong, Z., Khandoker, A. H., Hadjileontiadis, L. J., Leeson, P., & Lapidaire, W. (2023). The role of artificial intelligence in hypertensive disorders of pregnancy: towards personalized healthcare. Expert Review of Cardiovascular Therapy, 21(7), 531-543. https://www.tandfonline.com/doi/abs/10.1080/14779072.2023.2223978

[4] Bender, B. F., & Berry, J. A. (2023). Trends in passive IoT biomarker monitoring and machine learning for cardiovascular disease management in the US elderly population. Advances in geriatric medicine and research, 5(1), e230002. https://pmc.ncbi.nlm.nih.gov/articles/PMC10237513/

[5] Burke, J., Akbari, A., Bailey, R., Fasusi, K., Lyons, R. A., Pearson, J., ... & Schofield, D. (2023). Representing multimorbid disease progressions using directed hypergraphs. medRxiv, 2023-08. https://www.medrxiv.org/content/10.1101/2023.08.31.23294903v2

[6] Chakilam, C. (2022). Integrating Machine Learning and Big Data Analytics to Transform Patient Outcomes in Chronic Disease Management. *Journal of Survey in Fisheries Sciences*, 9(3), 118-130.

[7] Chushig M, C. D. (2022). Unsupervised Learning Methods to Extract ClinicalKnowledge of Patients with Chronic Diseases.

[8] Dataset Link: (n.d) https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[9] Eskofier, B. M., & Klucken, J. (2023). Predictive models for health deterioration: Understanding disease pathways for personalized medicine. Annual Review of Biomedical Engineering, 25(1), 131-156. https://www.annualreviews.org/content/journals/10.1146/annurev-bioeng-110220-030247

[10] Gokuladhas, S., Zaied, R. E., Schierding, W., Farrow, S., Fadason, T., & O'Sullivan, J. M. (2022). Integrating Multimorbidity into a Whole-Body Understanding of Disease Using Spatial Genomics. In Nuclear, Chromosomal, and Genomic Architecture in Biology and Medicine (pp. 157-187). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-06573-6_5

[11] Gokuladhas, S., Zaied, R. E., Schierding, W., Farrow, S., Fadason, T., & O'Sullivan, J. M. (2022). Integrating Multimorbidity into a Whole-Body Understanding of Disease Using Spatial Genomics. In Nuclear, Chromosomal, and Genomic Architecture in Biology and Medicine (pp. 157-187). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-06573-6_5

[12] Guasti, L., Dilaveris, P., Mamas, M. A., Richter, D., Christodorescu, R., Lumens, J., ... & Cowie, M. R. (2022). Digital health in older adults for the prevention and management of cardiovascular diseases and frailty. A clinical consensus statement from the ESC Council for Cardiology Practice/Taskforce on Geriatric Cardiology, the ESC Digital Health Committee and the ESC Working Group on e-Cardiology. ESC heart failure, 9(5), 2808-2822. https://onlinelibrary.wiley.com/doi/full/10.1002/ehf2.14022

[13] Gunderson, E. P., Greenberg, M., Sun, B., Goler, N., Go, A. S., Roberts, J. M., ... & Alexeeff, S. E. (2023). Early pregnancy systolic blood pressure patterns predict early-and later-onset preeclampsia and gestational hypertension among ostensibly low-to-moderate risk groups. *Journal of the American Heart Association, 12(*15), e029617. https://www.ahajournals.org/doi/full/10.1161/JAHA.123.029617

[14] Hanke, S., Mangialasche, F., Bödenler, M., Neumayer, B., Ngandu, T., Mecocci, P., ... & Stögmann, E. (2022). AI-based predictive modelling of the onset and progression of dementia. Smart Cities, 5(2), 700-714. https://www.mdpi.com/2624-6511/5/2/36

[15] Hasan, M. E., Islam, M. J., Islam, M. R., Chen, D., Sanin, C., & Xu, G. (2023). Applications of artificial intelligence for health informatics: A systematic review. J Artif Intell Med Sci, 4(2), 19-46. https://file.aisacademy.org.cn/journal/article/26/JOAIMS-4-2-19.pdf

[16] Hassaine, A., Salimi-Khorshidi, G., Canoy, D., & Rahimi, K. (2020). Untangling the complexity of multimorbidity with machine learning. Mechanisms of ageing and development, 190, 111325. https://www.sciencedirect.com/science/article/pii/S0047637420301214

[17] Hautala, A. J., Shavazipour, B., Afsar, B., Tulppo, M. P., & Miettinen, K. (2023). Machine learning models in predicting health care costs in patients with a recent acute coronary syndrome: a prospective pilot study. *Cardiovascular Digital Health Journal,* 4(4), 137-142. https://www.sciencedirect.com/science/article/pii/S2666693623000300

[18] Holm, N. N., Frølich, A., Andersen, O., Juul-Larsen, H. G., & Stockmarr, A. (2023). Longitudinal models for the progression of disease portfolios in a nationwide chronic heart disease population. Plos one, 18(4), e0284496. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0284496

[19] Hsu, W., Warren, J. R., & Riddle, P. J. (2022). Medication adherence prediction through temporal modelling in cardiovascular disease management. BMC Medical Informatics and Decision Making, 22(1), 313. https://link.springer.com/article/10.1186/s12911-022-02052-9

[20] Jain, A., Charpignon, M. L., Chen, I. Y., Philippakis, A., & Alaa, A. (2023). Generating new drug repurposing hypotheses using disease-specific hypergraphs. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024 (pp. 261-275). https://www.worldscientific.com/doi/abs/10.1142/9789811286421_0021

[21] Jain, L. (2023). Artificial Intelligence and Machine Learning for Healthcare.https://www.researchgate.net/profile/Juan-Velasquez-3/publication/363926877_Artificial_Intelligence_for_the_Future_of_Medicine/links/645ba02a2edb8e5f094db365/Artificial-Intelligence-for-the-Future-of-Medicine.pdf

[22] Jędrzejczyk, M., Foryś, W., Czapla, M., & Uchmanowicz, I. (2022). Relationship between multimorbidity and disability in elderly patients with coexisting frailty syndrome. *International Journal of Environmental Research and Public Health,* 19(6), 3461. https://www.mdpi.com/1660-4601/19/6/3461

[23] Jiang, M., Tian, S., Lin, X., Belsky, D., Guo, X., Huang, T., ... & Gao, X. (2023). Accelerated biological aging, cardiovascular risks, and cardiometabolic multimorbidity: a cohort study. https://www.researchsquare.com/article/rs-2710607/v1

[24] Kaur, N. (2025). Cardiovascular disease in Type 2 Diabetes Mellitus: A precision medicine approach applying artificial intelligence for heart failure and mortality prediction (Doctoral dissertation, University of Glasgow). https://theses.gla.ac.uk/85292/

[25] Khan, S. S., Coresh, J., Pencina, M. J., Ndumele, C. E., Rangaswami, J., Chow, S. L., ... & American Heart Association. (2023). Novel prediction equations for absolute risk assessment of total cardiovascular disease incorporating cardiovascular-kidney-metabolic health: a scientific statement from the American Heart Association. Circulation, 148(24), 1982-2004. https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000001191

[26] Kremer, R., Raza, S. M., Eto, F., Casement, J., Atallah, C., Finer, S., ... & Missier, P. (2022, December). Tracking trajectories of multiple long-term conditions using dynamic patient-cluster associations. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 4390-4399). IEEE. https://ieeexplore.ieee.org/abstract/document/10021034

[27] Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of ambient intelligence and humanized computing, 14(7), 8459-8486. https://link.springer.com/article/10.1007/S12652-021-03612-Z

[28] Li, C., Pan, Y., Zhang, R., Huang, Z., Li, D., Han, Y., ... & Kelly, T. N. (2023). Genomic innovation in early life cardiovascular disease prevention and treatment. Circulation research, 132(12), 1628-1647. https://www.thelancet.com/journals/landia/article/PIIS2213-8587(23)00086-4/abstract

[29] Li, S., Wang, Z., Vieira, L. A., Zheutlin, A. B., Ru, B., Schadt, E., ... & Li, L. (2022). Improving preeclampsia risk prediction by modeling pregnancy trajectories from routinely collected electronic medical record data. npj Digital Medicine, 5(1), 68. https://www.nature.com/articles/s41746-022-00612-x

[30] Liu, Q., Yu, M., Xiong, B., Cai, Z., Zhang, P., & Tan, C. W. (2023). Health analytics in business research: a literature review. *Journal of Management Analytics*, 10(3), 447-476. https://www.tandfonline.com/doi/abs/10.1080/23270012.2023.2209863

[31] Ly, B., Pop, M., Cochet, H., Duchateau, N., O'regan, D., & Sermesant, M. (2023). Outcome Prediction. In AI and Big Data in Cardiology: A Practical Guide (pp. 105-133). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-05071-8_6

[32] Maddula, R., MacLeod, J., McLeish, T., Painter, S., Steward, A., Berman, G., ... & Brown, S. A. (2022). The role of digital health in the cardiovascular learning healthcare system. Frontiers in cardiovascular medicine, 9, 1008575. https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2022.1008575/full

[33] Mbata, A. O., Soyege, O. S., Nwokedi, C. N., Tomoh, B. O., Mustapha, A. Y., Balogun, O. D., ... & Iguma, D. R. (2024). Preventative medicine and chronic disease management: reducing healthcare costs and improving long-term public health. *International Journal of Multidisciplinary Research and Growth Evaluation,* 5(06), 1584-1600. https://www.researchgate.net/profile/Olakunle-Soyege/publication/390694904_Preventative_Medicine_and_Chronic_Disease_Management_Reducing_Healthcare_Costs_and_Improving_Long-Term_Public_Health/links/67fb28eadf0e3f544f4102e5/Preventative-Medicine-and-Chronic-Disease-Management-Reducing-Healthcare-Costs-and-Improving-Long-Term-Public-Health.pdf

[34] Nuka, S. T. (2020). Predictive Modeling in Healthcare: Early Diagnosis and Patient Risk Profiling Using Machine Learning. Global Research Development (GRD) ISSN: 2455-5703, 5(12), 96-115. https://grdjournals.com/index.php/grdjournal/article/view/136

[35] Oikonomou, E. K., & Khera, R. (2023). Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovascular Diabetology, 22(1), 259. https://link.springer.com/article/10.1186/s12933-023-01985-3

[36] Qiu, L., Gorantla, S., Rajan, V., & Tan, B. C. (2021). Multi-disease predictive analytics: A clinical knowledge-aware approach. ACM Transactions on Management Information Systems (TMIS), 12(3), 1-34. https://dl.acm.org/doi/abs/10.1145/3447942

[37] Ruiz, R. B., & Velásquez, J. D. (2022). Artificial intelligence for the future of medicine. In Artificial Intelligence and Machine Learning for Healthcare: Vol. 2: Emerging Methodologies and Trends (pp. 1-28). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-11170-9_1

[38] Saha, P. K., Nadeem, S. A., & Comellas, A. P. (2023). A survey on artificial intelligence in pulmonary imaging. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13(6), e1510. https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1510

[39] Salah, H., & Srinivas, S. (2022). Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. Scientific Reports, 12(1), 21905. https://www.nature.com/articles/s41598-022-25933-5

[40] Shafqat, S., Anwar, Z., Rasool, R. U., Javaid, Q., & Ahmad, H. F. (2023). Rules extraction, diagnoses and prognosis of diabetes and its comorbidities using deep learning analytics with semantics on big data. Qeios, volume (issue), page numbers.

[41] Singer, M., & Mendenhall, E. (2022). Syndemics in global health. A companion to medical anthropology, 126-144. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119718963.ch7

[42] Srinivasan, K., & Jiang, J. (2023). Examining disease multimorbidity in US hospital visits before and during COVID-19 pandemic: a graph analytics approach. ACM Transactions on Management Information Systems, 14(2), 1-17. https://dl.acm.org/doi/full/10.1145/3564274

[43] Stefan, N., & Schulze, M. B. (2023). Metabolic health and cardiometabolic risk clusters: implications for prediction, prevention, and treatment. The lancet Diabetes & endocrinology, 11(6), 426-440. https://www.thelancet.com/journals/landia/article/PIIS2213-8587(23)00086-4/abstract

[44] Wang, Y., Wan, E. Y. F., Mak, I. L., Ho, M. K., Chin, W. Y., Yu, E. Y. T., & Lam, C. L. K. (2022). The association between trajectories of risk factors and risk of cardiovascular disease or mortality among patients with diabetes or hypertension: a systematic review. Plos one, 17(1), e0262885. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262885

[45] Wen, J., Nasrallah, I. M., Abdulkadir, A., Satterthwaite, T. D., Yang, Z., Erus, G., ... & AI4AD consortium. (2022). Novel genomic loci influence patterns of structural covariance in the human brain. medRxiv, 2022-07. https://www.medrxiv.org/content/10.1101/2022.07.20.22277727v2

[46] Wu, Y., Jia, M., Xiang, C., & Fang, Y. (2022). Latent trajectories of frailty and risk prediction models among geriatric community dwellers: an interpretable machine learning perspective. BMC geriatrics, 22(1), 900. https://link.springer.com/article/10.1186/s12877-022-03576-5

[47] Yang, L. (2023). Artificial Intelligence Methods for Discovery in Large Biobanks. Stanford University. https://www.proquest.com/openview/bac0a7cbc5f3b51bec6eb6554afbde64/1?pq-origsite=gscholar&cbl=18750&diss=y

[48] You, J., Guo, Y., Zhang, Y., Kang, J. J., Wang, L. B., Feng, J. F., ... & Yu, J. T. (2023). Plasma proteomic profiles predict individual future health risk. Nature communications, 14(1), 7817. https://www.nature.com/articles/s41467-023-43575-7

[49] ZEHRAOUI, F., Sendi, N., & Abchiche-Mimouni, N. (2022). MS-LSTMEA: Predicting clinical events for Hypertension using Multi-Sources LSTM Explainable Approach. Available at SSRN 4123459. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4123459