TechSaksham | edunet foundation

Medical Diagnosis using AI

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Kamepalli Narendra ,

nari73092@gmail.com

Under the Guidance of

Aditya Prasanth Ardak,

Master trainer,edunet foundation

ACKNOWLEDGEMENT

I would like to take this opportunity to express my deep sense of gratitude to all individuals who have helped me directly or indirectly throughout this thesis work.

Firstly, I would like to extend my heartfelt thanks to my supervisor, **Mr. Aditya Prasanth Ardak**, for being an exceptional mentor and advisor. His invaluable guidance, encouragement, and constructive criticism have been a constant source of inspiration and innovative ideas, significantly contributing to the successful completion of this project.

The confidence he placed in me served as a tremendous motivation throughout my journey. Working under his guidance over the past year has been a privilege. His unwavering support extended beyond project work to various aspects of the program, helping me grow as a responsible and skilled professional.

I am sincerely grateful for his insightful advice, lessons, and dedication, which have played a crucial role in shaping the success of this work.

ABSTRACT

This project presents an **AI-based medical diagnosis system** developed using **Support Vector Machine (SVM)** algorithms to accurately predict various diseases, including **Heart Disease, Lung Cancer, Parkinson's Disease, and Thyroid Disorders**. The primary objective is to provide a reliable diagnostic tool that assists healthcare professionals by delivering accurate predictions based on patient data.

Datasets for each disease are pre-processed, and essential features are extracted to optimize model performance. SVC and Logistic regression, known for its efficiency in classification problems, is employed to construct predictive models for each condition. The models are rigorously trained and validated to achieve high accuracy, ensuring precise and dependable diagnosis.

A **Streamlit application** is implemented to offer a simple, interactive, and user-friendly interface where users can input medical data and obtain diagnostic predictions in real-time. This platform aims to facilitate early detection and timely medical intervention by providing rapid diagnostic insights.

The proposed system showcases the potential of AI in enhancing healthcare diagnostics by offering accessible, scalable, and efficient tools. Future enhancements will focus on improving model accuracy, expanding the range of detectable diseases, and refining the user interface for better usability.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1Problem Statement:

The early and accurate diagnosis of critical diseases such as Heart Disease, Lung Cancer, Parkinson's Disease, and Thyroid Disorders remains a significant challenge in the healthcare industry. Delayed or inaccurate diagnoses often result in ineffective treatment plans, negatively impacting patient outcomes and increasing mortality rates. Medical professionals are under constant pressure to make swift, precise diagnoses based on various parameters, often relying on manual processes that can be time-consuming and prone to errors.

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) provides an opportunity to enhance the diagnostic process by offering automated, efficient, and reliable prediction systems. However, developing a generalized AI-based diagnostic tool that effectively handles multiple diseases while maintaining high accuracy and usability remains a complex task. Most existing models focus on individual diseases, lacking a comprehensive approach to address diverse medical conditions within a unified framework. Additionally, challenges related to data preprocessing, feature selection, model training, and real-time prediction must be addressed to ensure robustness and efficiency.

This project aims to build an AI-based medical diagnosis system using Support Vector Classifier (SVC) and Logistic Regression algorithms to predict the likelihood of various diseases accurately. The system will be designed to assist healthcare professionals by providing rapid diagnostic predictions based on patient data. A user-friendly Streamlit application will be developed to offer real-time predictions, enhancing accessibility and usability. By addressing the limitations of existing diagnostic methods and integrating multiple disease prediction models into a single platform, this system strives to improve diagnostic accuracy and support timely medical intervention.

1.2Motivation:

The motivation behind this project arises from the growing need for early and accurate disease detection in the medical field. Many life-threatening diseases such as Heart Disease, Lung Cancer, Parkinson's Disease, and Thyroid Disorders often go undiagnosed or are detected at advanced stages due to the lack of accessible diagnostic tools, leading to poor patient outcomes and increased mortality rates. The conventional diagnostic methods are often time-consuming, expensive, and require expert interpretation, which may not be readily available in all healthcare settings.

The potential applications of this project are vast and impactful. By developing an AI-based medical diagnosis system using Support Vector Classifier (SVC) and Logistic Regression algorithms, we aim to provide a robust, scalable, and efficient tool that can assist healthcare professionals in making quick and reliable diagnoses. Such a system can be particularly beneficial for remote areas, telemedicine platforms, and preliminary screening tools, where access to specialized medical expertise is limited.

The successful implementation of this project will contribute to improving healthcare accessibility and efficiency. It will empower healthcare providers with a powerful decision-support tool, enhancing diagnostic accuracy and enabling timely medical intervention. Additionally, the Streamlit application will ensure ease of use, making it accessible to a broader audience. Ultimately, this project strives to demonstrate the potential of AI in revolutionizing the healthcare sector by offering cost-effective, user-friendly, and high-precision diagnostic solutions.

1.3Objective:

The primary objective of this project is to develop an AI-based medical diagnosis system capable of accurately predicting multiple diseases, including Heart Disease, Lung Cancer, Parkinson's Disease, and Thyroid Disorders, using Support Vector Classifier (SVC) and Logistic regression algorithms. The system aims to provide healthcare professionals and patients with a reliable tool for early detection and diagnosis, enhancing medical decision-making and treatment planning.

The specific objectives of this project are:

1. To build predictive models using SVM algorithms for each targeted disease, ensuring high accuracy, robustness, and efficiency.

2. To preprocess and analyze medical datasets, extracting relevant features to enhance model performance and reliability.

3. To integrate all predictive models into a unified platform accessible via a user-friendly Streamlit application for real-time diagnostic predictions.

4. To evaluate and validate the models through rigorous testing, ensuring accurate predictions across various disease datasets.

5. To provide a scalable, efficient, and accessible diagnostic tool that can potentially assist healthcare professionals and patients, particularly in remote or underserved areas.

6. To improve healthcare accessibility and efficiency by demonstrating the feasibility of AI-driven diagnostic systems in clinical settings.

1.4 Scope of the Project:

The scope of this project is centered around developing a robust AI-based medical diagnosis system using Support Vector Classifier (SVC) and Logistic regression algorithms for the prediction of multiple diseases, namely Heart Disease, Lung Cancer, Parkinson's Disease, and Thyroid Disorders. The project aims to build accurate predictive models by preprocessing medical datasets, extracting relevant features, and implementing efficient training techniques. The primary goal is to achieve high-performance metrics, including precision, recall, F1-score, and accuracy, ensuring reliable diagnostic predictions.

Furthermore, this project emphasizes the creation of a user-friendly interface using Streamlit. The interactive web application will allow users to input medical data and obtain real-time diagnostic predictions. By integrating all disease-specific models into a single platform, the system aims to provide an accessible and efficient tool for medical diagnosis. The deployment of this platform will focus on ensuring compatibility and seamless interaction between the backend models and the frontend application, enhancing the user experience.

The project also includes rigorous evaluation and validation of the predictive models using appropriate datasets. Techniques such as hyperparameter tuning and model optimization will be employed to improve prediction accuracy and robustness. The system will be thoroughly tested to ensure reliability and effectiveness in various scenarios.

CHAPTER 2

Literature Survey

The term diagnosis is used for finding symptoms of disease or analysis of the patients to determine the health conditions. The diagnosis is usually performed through one of these methods, i.e., examining the physical condition of the patient, exploring patient's history, or from diagnostic tests which are analysed by various healthcare professionals such as dentists, physicians, chiropractors, physical therapists, or physician assistants and compounders etc. [1]. The patient's history is frequently saved in the form of a prescription for necessary medications, streamline workflow, and to keep track of the patient's performance. Initially, the prescription was saved in the form of the paper chart containing the type of diseases, suggested medicines, vaccination dates, treatment plans, and the test results of X-rays specific hospitals. However, in the modern age of the computer, the prescription is saved in a digital format which is known as an electronic medical record (EMR) or electronic health record (EHR). These electronic records help the physicians to access the patient's records instantly, to keep track of patients' due dates for checkups and immunizations and monitor patients' health performance and make decisions accordingly [2].

Although both these terms (EMR and EHR) are used interchangeably, according to the 'Office of the National Coordinator of Health Information Technology (ONC)' both the terms are utilized exclusively [3]. The EMR is the digital form of the prescription, which contains the patients' information collected in a provider's office for healthcare professionals. The EMR data can be either human-generated or machine-generated [4]. EMRs have multiple advantages over paper prescriptions including instant access, keeping track of patients' information, saving patients visits, screening patients, and enhancing the healthcare's quality [5]. The scope of the EHRs goes beyond EMRs, as it contains information of all the medical investigators involved in patients' health records. The patient's information is also shared with other clinicians and medical researchers in various hospitals to study and improve the root causes of the disease. EMRs contain temporal and heterogeneous doctor order information which may be used as an input for treatment pattern discovery [6].

EHRs also facilitate the patients to see their records on how the progress is going on, which motivates them in many cases (not necessarily) [7], [8]. Though EMR and EHR provide many benefits to users and practitioners, there are many challenges associated with the implementation of these electronic records including downtime of the computers, incapability of the computer experts, lack of communication among users, and security threats of confidentiality leakage [9]–[11].

While existing electronic medical record (EMR) and electronic health record (EHR) systems have made significant advancements in healthcare management, several limitations and gaps remain. First, many existing systems primarily focus on data storage and retrieval, lacking advanced analytical capabilities to predict or diagnose diseases accurately. These systems often rely on rule-based approaches, which may not generalize well across diverse patient populations.

Secondly, existing solutions frequently face challenges related to interoperability, where patient records from different healthcare providers or institutions are not easily integrated, leading to incomplete or fragmented information. Additionally, privacy and security concerns remain prevalent, as breaches or unauthorized access to sensitive patient data can compromise confidentiality.

Furthermore, current diagnostic systems often require extensive manual input and human expertise, making them time-consuming and susceptible to human errors. The use of traditional models without advanced machine learning algorithms may also limit the accuracy of predictions and diagnostic capabilities.

Our project aims to address these limitations by developing an AI-based medical diagnosis system using Support Vector Classifier (SVC) and Logistic regression models through Streamlit. This system is designed to provide accurate and efficient predictions based on patient data, enhancing diagnostic precision and reducing the need for manual intervention. The integration of AI ensures the model's ability to generalize across diverse datasets, improving its applicability to various healthcare scenarios. Additionally, the user-friendly interface built with Streamlit offers seamless interaction, making it accessible to medical professionals and patients alike.

CHAPTER 3

Proposed Methodology

3.1 System Design

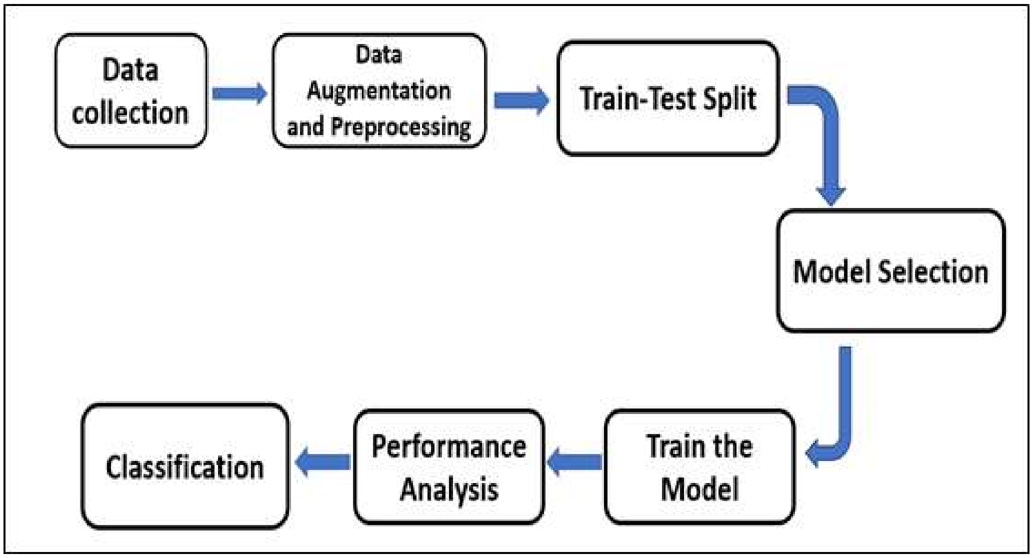
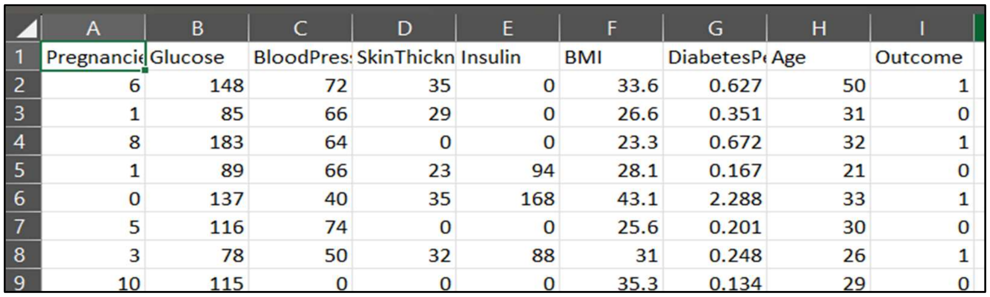


Fig 1: Methodology Diagram

Data Collection

The data collection process involves gathering relevant datasets from various reliable sources to build an effective disease detection system. For this project, datasets for multiple diseases are collected, each containing a set of features essential for accurate prediction and diagnosis. The datasets include:

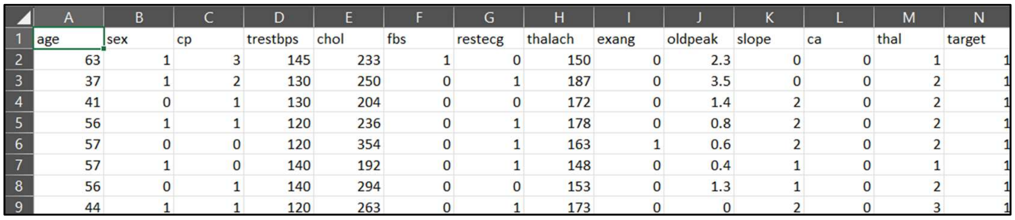
1. **Diabetes Dataset:**
   - Features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome.
   - Description: This dataset is used to predict the presence of diabetes in patients based on various health metrics.



	A	B	C	D	E	F	G	H	I
1	Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0

Fig 2: Diabetes Dataset

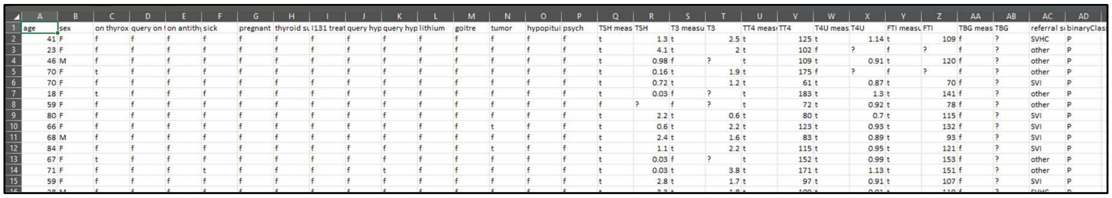
2. **Heart Disease Dataset:**
   - Features: Age, Sex, CP (Chest Pain Type), Trestbps (Resting Blood Pressure), Chol (Serum Cholesterol), FBS (Fasting Blood Sugar), Restecg (Resting Electrocardiographic Results), Thalach (Maximum Heart Rate), Exang (Exercise Induced Angina), Oldpeak (ST Depression), Slope, CA (Number of Major Vessels), Thal, Target.
   - Description: This dataset helps predict the presence of heart disease by analyzing patient health indicators.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1

Fig 3: Heart Disease Dataset

3. **Hypothyroid Dataset:**
   - Features: Age, Sex, On Thyroxine, Query on Thyroxine, On Antithyroid Medication, Sick, Pregnant, Thyroid Surgery, I131 Treatment, Query Hypothyroid, Query Hyperthyroid, Lithium, Goitre, Tumor, Hypopituitary, Psych, TSH Measured, TSH, T3 Measured, T3, TT4 Measured, TT4, T4U Measured, T4U, FTI Measured, FTI, TBG Measured, TBG, Referral Source, BinaryClass.
   - Description: This dataset is used to classify patients as hypothyroid or non-hypothyroid based on various medical attributes.

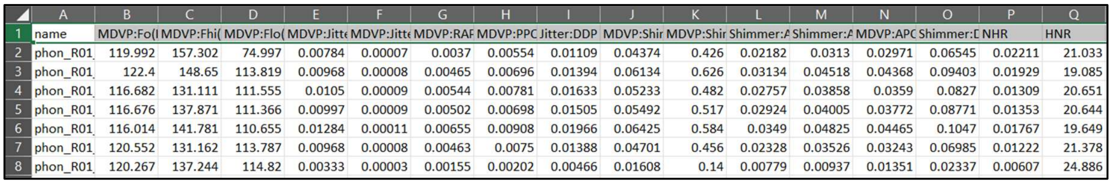


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	age	sex	on thyrox	query on t	on antithy	sick	pregnant	thyroid su	I131 treat	query hyp	query hyp	lithium	goitre	tumor	hypopitui	psych	TSH meas	TSH	T3 measu	T3	TT4 meas	TT4	T4U meas	T4U	FTI measu	FTI	TBG meas	TBG	referral s	binaryClas
2	41	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.3	t	2.5	t	125	t	1.14	t	109	f	?	SVHC	P
3	23	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	4.1	t	2	t	102	f	?	f	?	f	?	other	P
4	46	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.98	f	?	t	109	t	0.91	t	120	f	?	other	P
5	70	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.16	t	1.9	t	175	f	?	f	?	f	?	other	P
6	70	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.72	t	1.2	t	61	t	0.87	t	70	f	?	SVI	P
7	18	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.03	f	?	t	183	t	1.3	t	141	f	?	other	P
8	59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	?	f	?	t	72	t	0.92	t	78	f	?	other	P
9	80	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	2.2	t	0.6	t	80	t	0.7	t	115	f	?	SVI	P
10	66	F	f	f	f	f	f	f	f	f	f	f	f	t	f	f	t	0.6	t	2.2	t	123	t	0.93	t	132	f	?	SVI	P
11	68	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	2.4	t	1.6	t	83	t	0.89	t	93	f	?	SVI	P
12	84	F	f	f	f	f	f	f	f	f	f	f	f	t	f	f	t	1.1	t	2.2	t	115	t	0.95	t	121	f	?	SVI	P
13	67	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.03	f	?	t	152	t	0.99	t	153	f	?	other	P
14	71	F	f	f	f	t	f	f	f	f	t	f	f	f	f	f	t	0.03	t	3.8	t	171	t	1.13	t	151	f	?	other	P
15	59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	2.8	t	1.7	t	97	t	0.91	t	107	f	?	SVI	P

Fig 4 : Hyperthyroid Dataset

4. **Parkinson's Disease Dataset:**
   - Features: Name, MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, NHR, HNR, Status, RPDE, DFA, Spread1, Spread2, D2, PPE.

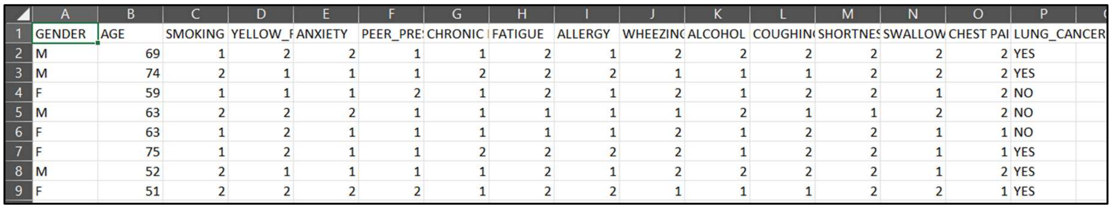
- Description: This dataset helps in detecting Parkinson’s Disease by analyzing vocal features.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	name	MDVP:Fo(I	MDVP:Fhi(	MDVP:Flo(	MDVP:Jitte	MDVP:Jitte	MDVP:RAF	MDVP:PPC	Jitter:DDP	MDVP:Shir	MDVP:Shir	Shimmer:A	Shimmer:A	MDVP:APC	Shimmer:D	NHR	HNR
2	phon_R01	119.992	157.302	74.997	0.00784	0.00007	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033
3	phon_R01	122.4	148.65	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	0.626	0.03134	0.04518	0.04368	0.09403	0.01929	19.085
4	phon_R01	116.682	131.111	111.555	0.0105	0.00009	0.00544	0.00781	0.01633	0.05233	0.482	0.02757	0.03858	0.0359	0.0827	0.01309	20.651
5	phon_R01	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	0.517	0.02924	0.04005	0.03772	0.08771	0.01353	20.644
6	phon_R01	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	0.584	0.0349	0.04825	0.04465	0.1047	0.01767	19.649
7	phon_R01	120.552	131.162	113.787	0.00968	0.00008	0.00463	0.0075	0.01388	0.04701	0.456	0.02328	0.03526	0.03243	0.06985	0.01222	21.378
8	phon_R01	120.267	137.244	114.82	0.00333	0.00003	0.00155	0.00202	0.00466	0.01608	0.14	0.00779	0.00937	0.01351	0.02337	0.00607	24.886

Fig 5: Parkinson’s Disease Dataset

5. **Lung Cancer Dataset:**
   - Features: Gender, Age, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, Chest Pain, Lung Cancer.
   - Description: This dataset helps predict lung cancer presence based on lifestyle habits and symptoms.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	GENDER	AGE	SMOKING	YELLOW_F	ANXIETY	PEER_PRE	CHRONIC	FATIGUE	ALLERGY	WHEEZING	ALCOHOL	COUGHIN	SHORTNES	SWALLOW	CHEST PAI	LUNG_CANCER
2	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
3	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
4	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
5	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
6	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
7	F	75	1	2	1	1	2	2	2	2	1	2	2	1	1	YES
8	M	52	2	1	1	1	1	2	1	2	2	2	2	1	2	YES
9	F	51	2	2	2	2	1	2	2	1	1	1	2	2	1	YES

Fig 6 : Lung Cancer Dataset

The data is collected in CSV format for easy loading and processing. Each dataset will be analysed and pre-processed to remove noise, handle missing values, and normalize features for optimal model training.

Data Preprocessing

Data preprocessing is a critical step to ensure the datasets are properly prepared for training the machine learning models. It involves cleaning, transforming, and organizing the data to enhance model performance and accuracy. The preprocessing steps include:

1. **Loading Data:**
   - Load the CSV files for each disease dataset using pandas.
2. **Handling Missing Values:**
   - Identify missing or null values in the dataset.
   - Use techniques such as mean/mode/median imputation or deletion of records if necessary.

3. **Encoding Categorical Data:**
    - Convert categorical variables to numerical representations using techniques like:
        - Label Encoding for binary categorical variables.
        - One-Hot Encoding for multi-class categorical variables.
4. **Feature Scaling:**
    - Normalize or standardize features to ensure uniformity in data distribution, especially when models are sensitive to feature scales (e.g., SVM, Neural Networks).
5. **Removing Irrelevant Features:**
    - Drop unnecessary columns that do not contribute to the prediction process (e.g., 'Name' in Parkinson's dataset).
6. **Data Splitting:**
    - Split the data into training and testing sets using techniques like:
        - Train-Test Split (e.g., 80% training, 20% testing).
        - Cross-Validation for better generalization.
7. **Data Balancing:**
    - Address class imbalance using techniques such as:
        - SMOTE (Synthetic Minority Over-sampling Technique).
        - Undersampling the majority class.
8. **Feature Selection:**
    - Identify and select the most relevant features contributing to predictions using techniques like:
        - Correlation Matrix Analysis.
        - Recursive Feature Elimination (RFE).
        - Principal Component Analysis (PCA) if necessary.
9. **Data Augmentation (if applicable):**
    - Generate new samples from the existing dataset to improve model robustness (e.g., for image datasets).
10. **Data Transformation (if applicable):**

- Apply log transformations, polynomial features, or other transformations to improve model performance.

  The preprocessed data is now ready for training and testing the models.

Train-Test Split

The train-test split is a crucial step in building machine learning models, where the dataset is divided into two parts: a training set and a testing set. The training set is used to train the model, enabling it to learn patterns and relationships within the data. Typically, 70-80% of the dataset is allocated for training. The testing set, comprising the remaining 20-30%, is then used to evaluate the model's performance by making predictions on unseen data. This process helps in assessing the model's generalization ability and ensures that the model is not overfitting to the training data. To maintain reliability, it's essential to perform random shuffling of data before splitting, and techniques like K-Fold Cross-Validation can further enhance evaluation by providing a more robust measure of model performance.

Model Selection

In this step, appropriate machine learning models are selected based on the type of disease being predicted and the nature of the dataset. For each disease, different models are chosen considering their suitability for classification tasks. The models used for each disease are:

Disease	**Model**
Heart Disease	Logistic Regression
Lung Cancer	Logistic Regression
Parkinson's Disease	Support Vector Classifier
Thyroid	Logistic Regression
Diabetes	Logistic Regression

Table 1: Disease and Models used

Logistic Regression is chosen for most diseases due to its simplicity, efficiency, and robustness in binary classification tasks. For Parkinson's Disease, a Support Vector Classifier is preferred as it can handle high-dimensional data effectively and is suitable for non-linear classification problems. The selected models are trained using the training datasets, and their performance is evaluated using various metrics to ensure accuracy and reliability.

1. Logistic Regression :

Logistic Regression is a statistical method used for binary classification problems where the target variable has two possible outcomes, such as disease presence or absence. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability of a given data point belonging to a specific category by applying a logistic (sigmoid) function to the output. The logistic function maps the predicted values between 0 and 1, making it suitable for classification tasks. The model learns the relationship between the independent variables (features) and the dependent variable (outcome) by estimating coefficients through optimization techniques like gradient descent.

Logistic Regression is widely used due to its simplicity, efficiency, and interpretability. It is particularly effective when the data is linearly separable, offering high accuracy in distinguishing between two classes. Moreover, the model provides probabilities as outputs, which allows for threshold tuning to adjust sensitivity and specificity according to the problem's requirements. In medical diagnosis systems, Logistic Regression is commonly applied because it helps determine the likelihood of disease presence based on patient data. Additionally, it's computationally efficient and robust against overfitting when regularization techniques like L1 (Lasso) or L2 (Ridge) regularization are applied.

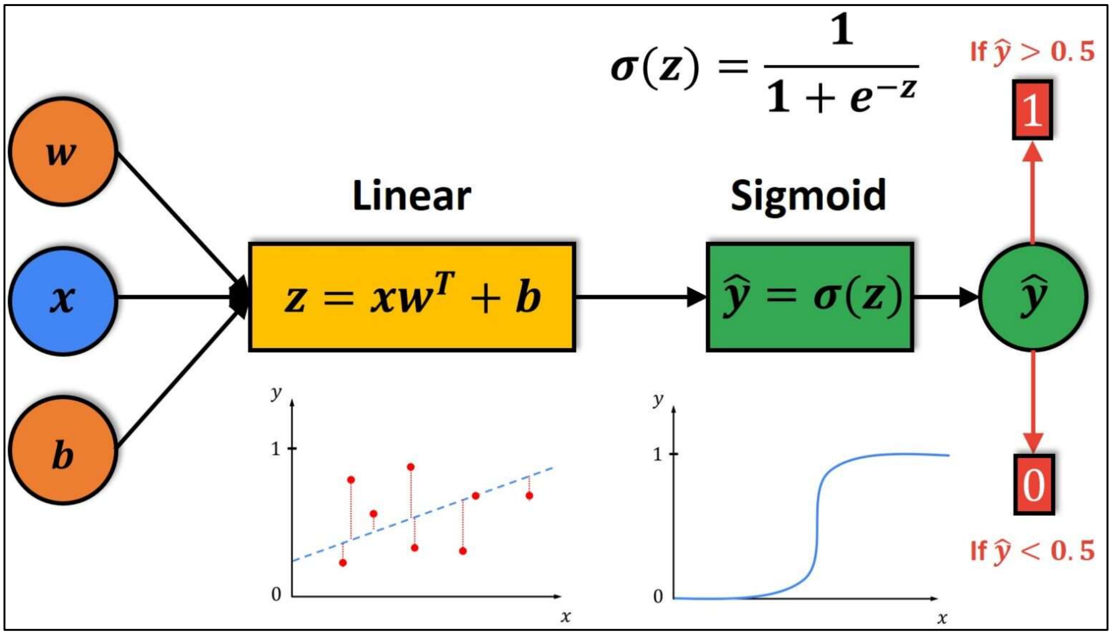


Fig 7: Logistic Regression Model

2. SVC:

Support Vector Classifier (SVC) is a supervised machine learning algorithm used for classification tasks. It is based on the concept of finding the hyperplane that best separates data points of different classes in a high-dimensional space. SVC is particularly effective for complex datasets where classes are not linearly separable by transforming the original feature space using kernel functions like linear, polynomial, radial basis function (RBF), and sigmoid. The RBF kernel is most commonly used for non-linear classification problems as it can model complex relationships between features.

SVC is known for its robustness and ability to handle high-dimensional data effectively. It performs well when there is a clear margin of separation between classes and is less prone to overfitting, especially in high-dimensional spaces. By maximizing the margin between classes, SVC achieves high generalization performance. In disease prediction systems, SVC is particularly useful for detecting conditions with complex patterns, such as Parkinson's Disease, where subtle variations in features can indicate the presence of disease. Its ability to work with non-linear boundaries makes it a powerful tool for various classification tasks.

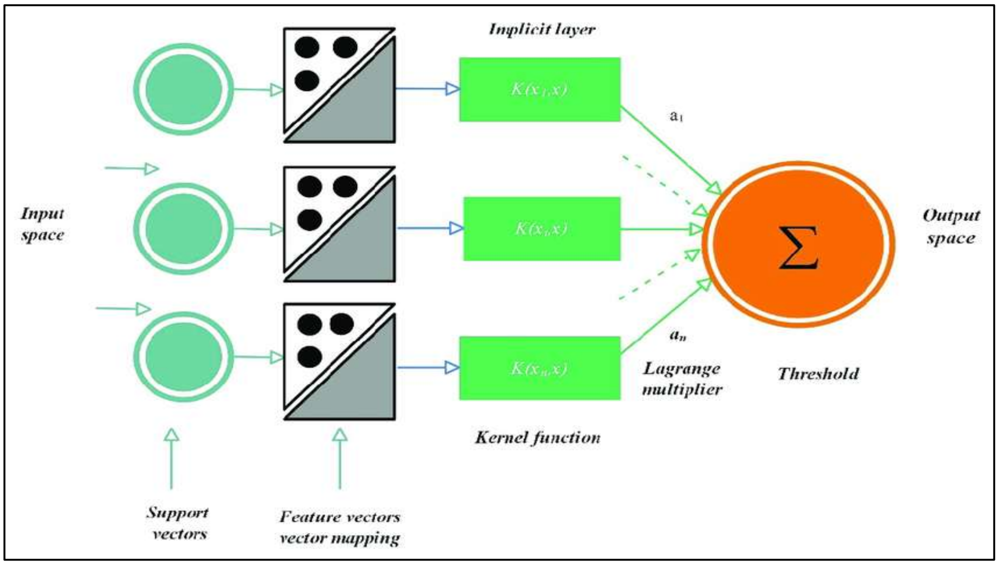


Fig 8: svc model

Model Training:

The training process for each disease prediction model involved preprocessing the data, splitting it into training and testing datasets, and fitting the models to the training data. For heart disease, lung cancer, thyroid, and diabetes prediction, Logistic Regression was chosen due to its simplicity, interpretability, and efficiency in handling binary classification tasks. On the other hand, Parkinson's disease detection utilized a Support Vector Classifier (SVC), which is well-suited for high-dimensional data and provides robust performance with appropriate kernel selection.

During training, each model was optimized to maximize accuracy while minimizing overfitting. This was achieved by tuning hyperparameters where necessary and ensuring proper data splitting. The models' performance was measured by comparing predictions on the testing dataset to the actual labels, calculating metrics such as accuracy, precision, and F1 scores. Consistent training and testing accuracies for most models indicate good generalization, suggesting that the models can effectively predict diseases on new, unseen data.

Performance Analysis:

The performance of various machine learning models applied to medical diagnosis tasks was evaluated based on their training and testing accuracies. For heart disease prediction, a Logistic Regression model achieved a training accuracy of 85.12% and a testing accuracy of 81.97%. This indicates that the model performs reasonably well, but there is a slight drop in accuracy during testing, suggesting possible overfitting or insufficient generalization to new data. The Logistic Regression model's simplicity and efficiency make it a suitable choice for heart disease prediction, where interpretability is crucial.

For lung cancer prediction, the Logistic Regression model exhibited high performance with both training and testing accuracies exceeding 93.5%. Such impressive results demonstrate that the model generalizes well to unseen data. This consistency between training and testing suggests that the features used are highly relevant, and the model is well-tuned for this particular classification task. Given the severity of lung cancer, a high-performing model like this is crucial for early and accurate diagnosis.

In the case of Parkinson's disease detection, a Support Vector Classifier achieved 87.18% accuracy for both training and testing, indicating strong generalization. The Thyroid disease prediction model, also built with Logistic Regression, demonstrated excellent performance, achieving approximately 95.6% accuracy for both training and testing. The robustness and reliability of these models suggest they can be effective tools in early detection and diagnosis of various diseases. However, further enhancement of the models through techniques like hyperparameter tuning and cross-validation could potentially boost performance even further.

Model Evaluation:

The evaluation of the trained models was carried out using metrics such as accuracy, precision, and F1 score. For most models, there was a strong alignment between training and testing accuracies, indicating effective generalization. The Logistic Regression models for heart disease, lung cancer, thyroid disease, and diabetes showed reliable performance with minimal overfitting, which confirms their suitability for binary classification tasks where interpretability and efficiency are important.

The Support Vector Classifier used for Parkinson's disease detection also demonstrated consistent accuracy across training and testing datasets, suggesting robustness in handling complex, high-dimensional data. While the models performed well overall, the evaluation process highlights potential areas for improvement, such as fine-tuning hyperparameters and incorporating feature selection techniques to further enhance model performance. Regular cross-validation could also contribute to improving generalization and boosting predictive accuracy.

Classification:

The classification models developed for predicting various diseases, including heart disease, lung cancer, Parkinson's disease, thyroid disease, and diabetes, are primarily binary classifiers, distinguishing between healthy and diseased states. Logistic Regression was selected for most diseases due to its efficiency in handling linearly separable data and its interpretability, making it suitable for clinical diagnosis where understanding feature impact is important. The Support Vector Classifier (SVC) was applied to Parkinson's disease detection, leveraging its ability to handle complex, high-dimensional data through kernel transformations.

During the classification process, the models were trained using labeled datasets and then tested on separate, unseen data to assess their generalization capabilities. The performance metrics collected—accuracy, precision, and F1 scores—were used to evaluate the reliability of each model. The models generally performed well, with training and testing accuracies being closely aligned, indicating minimal overfitting. However, improvements could be made by experimenting with alternative algorithms, fine-tuning hyperparameters, and enhancing feature engineering to boost classification performance.

3.2 Requirement Specification

To successfully implement the medical diagnosis models, several hardware and software resources are necessary. This section outlines the required tools and technologies.

3.2.1 Hardware Requirements:

- Processor: Intel Core i5 or above (Recommended: Intel Core i7/AMD Ryzen 5 or above)
- RAM: 8 GB or higher (Recommended: 16 GB or higher for faster processing)
- Storage: At least 10 GB of free disk space
- GPU: Optional (Recommended for deep learning models: NVIDIA GPU with CUDA support)

- Operating System: Windows 10/11, Linux (Ubuntu 20.04+), or macOS

3.2.2 Software Requirements:

- Programming Language: Python (Version 3.8 or above)
- Libraries/Frameworks:
  - Scikit-Learn (For model building and evaluation)
  - Pandas (For data manipulation and analysis)
  - NumPy (For numerical computations)
  - Matplotlib / Seaborn (For visualizations)
- Integrated Development Environment (IDE): Jupyter Notebook / VS Code / PyCharm
- Packages Installation: Pip or Conda
- Additional Tools: Pickle (For model serialization and deserialization)
- Dataset Format: CSV files for structured data input
- Version Control: Git (For version management)

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

Here are the metrics presented as separate tables for each model:

Heart Disease

Metric	Score
Training Accuracy	0.8512
Testing Accuracy	0.8197
Model	Logistic Regression

Table 2 : accuracy of Heart disease classifier

Lung Cancer

Metric	Score
Training Accuracy	0.9352
Testing Accuracy	0.9355
Model	Logistic Regression

Table 3 : accuracy of Lung Cancer classifier

Parkinson's Disease

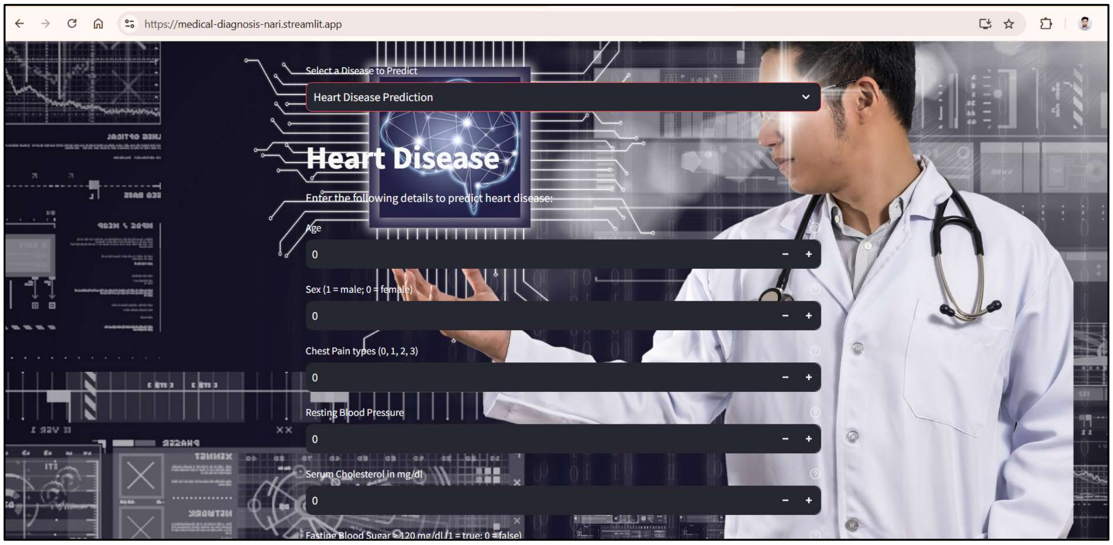
Metric	Score
Training Accuracy	0.8718
Testing Accuracy	0.8718
Model	Support Vector Classifier

Table 4: accuracy of Parkinson's Disease classifier

Thyroid

Metric	Score
Training Accuracy	0.9562
Testing Accuracy	0.9563
Model	Logistic Regression

Table 5: accuracy of heart disease classifier



https://medical-diagnosis-nari.streamlit.app

Select a Disease to Predict

Heart Disease Prediction

Heart Disease

Enter the following details to predict heart disease:

Age

0

Sex (1 = male; 0 = female)

0

Chest Pain types (0, 1, 2, 3)

0

Resting Blood Pressure

0

Serum Cholesterol in mg/dl

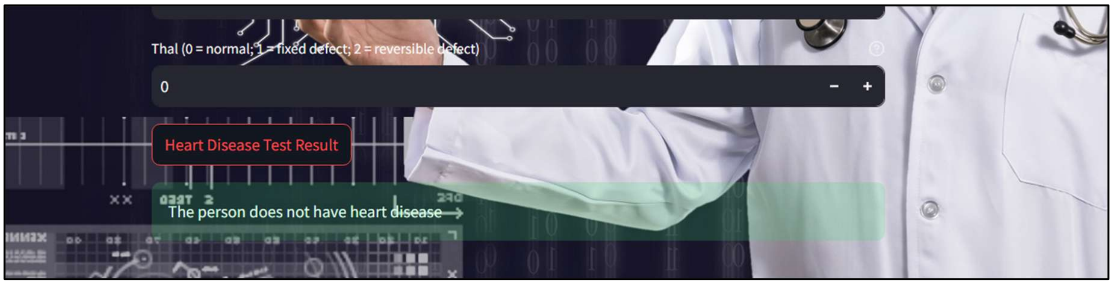
0

Fasting Blood Sugar > 120 mg/dl (1 = true; 0 = false)

Fig 9: medical imaging web site

The above image displays a visually appealing user interface of an AI-based medical diagnosis system for predicting diseases, specifically heart disease, developed using Streamlit. The background features a high-tech, futuristic theme with neural network visuals, enhancing the aesthetic appeal and conveying the purpose of the application. The interface allows the user to select a disease from a dropdown menu, with "Heart Disease Prediction" currently selected.

The form below the title gathers essential patient details required for prediction, including Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, and other relevant medical parameters. Each input field is represented by sleek slider controls or number input boxes, making it user-friendly and intuitive. The well-organized layout, complemented by a dark theme and glowing elements, offers a modern and sophisticated look. The design effectively combines practicality with a polished visual experience to enhance user engagement.



Thal (0 = normal; 1 = fixed defect; 2 = reversible defect)

0

The person does not have heart disease

Fig 10: result

The above image shows the result section of the AI-based medical diagnosis system for heart disease prediction. After the user enters all the required medical parameters, they can click the red button labeled "Heart Disease Test Result." Once the prediction is made, the result is displayed below the button in a green-colored text box with a soft glow effect, providing clear feedback to the user.

In this instance, the prediction states: “The person does not have heart disease”. The color scheme, with red for the button and green for the result, effectively conveys the message, enhancing the user experience by making the outcome immediately recognizable. The clean and modern design maintains consistency with the rest of the application.

Link : https://medical-diagnosis-nari.streamlit.app/

4.2 GitHub Link for Code:

https://github.com/Nari-2002/Medical-diagnosis-using-AI

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

In the future, several improvements can be made to enhance the accuracy and robustness of the AI-based medical diagnosis system. Firstly, expanding the dataset by including more diverse and comprehensive medical records can improve model generalization. Additionally, implementing advanced techniques such as ensemble learning and deep learning architectures may yield better predictive performance. Incorporating feature selection methods could also help identify the most critical parameters for each disease, thereby improving model efficiency. Moreover, integrating explainable AI techniques can provide more transparency and reliability to the predictions, making them more trustworthy for medical professionals. Finally, deploying the system with real-time monitoring capabilities and regularly updating it with new medical data will ensure the model remains accurate and relevant over time.

Conclusion:

The AI-based medical diagnosis system developed in this project demonstrates a promising approach to early and accurate disease prediction. By employing machine learning models such as Logistic Regression and Support Vector Classifier, the system effectively predicts multiple diseases, including Heart Disease, Lung Cancer, Parkinson's Disease, Thyroid Disease, and Diabetes, based on relevant health metrics. The user-friendly interface built using Streamlit provides a seamless experience for users to obtain rapid predictions. This project contributes to healthcare by offering a cost-effective, accessible, and efficient tool that can assist in preliminary medical diagnosis. Future enhancements, such as integrating more advanced models and expanding datasets, will further strengthen the system's reliability and applicability in real-world medical diagnostics.

REFERENCES

[1] M. Stewart, Patient-Centered Medicine: Transforming the Clinical Method. Oxford, U.K.: Radcliffe Publishing, 2003.

[2] J. Stausberg, D. Koch, J. Ingenerf, and M. Betzler, “Comparing paper-based with electronic patient records: Lessons learned during a study on diagnosis and procedure codes,” Journal of the American Medical Informatics Association, vol. 10, no. 5, pp. 470–477, Sep. 2003.

[3] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, “Challenges and opportunities of big data in health care: A systematic review,” JMIR Medical Informatics, vol. 4, no. 4, p. e38, Nov. 2016.

[4] J. J. Firthous and M. M. Sathik, “Survey on using electronic medical records (EMR) to identify the health conditions of the patients,” Journal of Engineering Science, vol. 11, no. 5, 2020.

[5] G. Makoul, R. H. Curry, and P. C. Tang, “The use of electronic medical records: Communication patterns in outpatient encounters,” Journal of the American Medical Informatics Association, vol. 8, no. 6, pp. 610–615, Nov. 2001.

[6] J. Chen, L. Sun, C. Guo, and Y. Xie, “A fusion framework to extract typical treatment patterns from electronic medical records,” Artificial Intelligence in Medicine, vol. 103, Mar. 2020, Art. no. 101782.

[7] S. Ajami and T. Bagheri Tadi, “Barriers for adopting electronic health records (EHRs) by physicians,” Acta Informatica Medica, vol. 21, no. 2, p. 129, 2013.

[8] S. U. Rehman, S. Tu, Y. Huang, and Z. Yang, “Face recognition: A novel unsupervised convolutional neural network method,” in Proceedings of the IEEE International Conference on Online Analysis and Computing Science (ICOACS), May 2016, pp. 139–144.

[9] W. R. Hersh, “The electronic medical record: Promises and problems,” Journal of the American Society for Information Science, vol. 46, no. 10, pp. 772–776, Dec. 1995.

[10] N. Jenkings and R. Wilson, “The challenge of electronic health records (EHRs) design and implementation: Responses of health workers to drawing a ‘big and rich picture’ of a future EHR programme using animated tools,” Journal of Innovation in Health Informatics, vol. 15, no. 2, pp. 93–101, Jun. 2007.

[11] M. Cifuentes, M. Davis, D. Fernald, R. Gunn, P. Dickinson, and D. J. Cohen, “Electronic health record challenges, workarounds, and solutions observed in practices integrating behavioral health and primary care,” Journal of the American Board of Family Medicine, vol. 28, no. 1, pp. S63–S72, Sep. 2015.