Journal of Theoretical and Applied Information Technology
31st January 2025. Vol.103. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TRANSFORMING IMAGES TO INSIGHTS: OCR-DRIVEN SENTIMENT ANALYSIS FOR MEDICAL DATACLASSIFICATION

G DATTA SAI SREYA[1], P.RAJARAJESWARI[2], S.HRUSHIKESAVA RAJU[3]

[1,2,3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India, 522302
dattasaisreya@gmail.com, rajilikhitha@kluniversity.in, hkesavaraju@gmail.com

ABSTRACT

The challenges of existing methods in the image to the text are time-consuming, lack of quality of images, and bringing to the format that is ready for analysis. To overcome these, and to provide better performance, Logistic regression, and Support Vector Machines are applied along with data augmentation for more accuracy. Optical Character Recognition is used to convert the written image into text format, in the review of all theexisting methodologies over data augmentation of sentiment analysis over classifying the input data as medical (or) non-medical classification would be time-consuming and intensive to differentiate and understand the real context behindthe data. We conducted a huge analysis of Medical and non-medical data during run-time our model will input thehigh-resolution image, and convert it to a text file from the input image, which is thereby processed by our proposed model to identify the accurate and fine-tuned algorithm. We have highlighted the significant approaches in NLP (Natural Language Processing). This classification between Medical and Non-Medical data provided prominent results with several pre-trained datasets, which resulted in fetching real-time information from medically approved web information.

Keywords: *Optical Character Recognition (OCR), Tesseract, Natural Language Processing (NLP), Logistic Regression, Support Vector Machine, Generative AI.*

I.INTRODUCTION

Classifying the data manually will be difficult for analyzing the data is medical or non-medical, for this, we have gone through training over a set of datasets, and predict the data accurately. For this optical character recognition performs well and is used, where it is embedded with tesseract is a popular type of extraction mechanism. Tesseract usage helps to identify the characters in an image and therefore studies all the lines and characters in the image, trains them, and predicts the output in a text file format. This research goes in two parts where one part recognizes the characters from the image and saves in a .txt format file. The second part is algorithm analysis to classify and train the model with a combination of two different ML algorithms. Logistic Regression is a much-supervised machine learning algorithm this predicts the data by a probability of an outcome, event or the observation. The confidence of the event based on its training and predicting probability of an outcome is rigours classified using logistic regression. This kind of algorithm model predicts a dependent data variable by analyzing its relationship between one or more existing independent variables. Support vector machine algorithm is the high accurate algorithm in our model by providing high accuracy than logistic regression. This algorithm goes hand in hand with classification and regression. It helps to solve a complex classifications, regression and outline detection with an optimal data transformation this will determine the output with predefined labels or classes. Based on Character recognition from the model using tesseract the data will be used as input in the next phase of model, the classification algorithms help in augmenting the data with labelling them by medical or non- medical classifications. Natural language processing helps us to interrogate the specific human text or image intoa predictive model of approach. Generative AI is a new method to classify or summarize the content that is provided as part of input, it debugs the data to analyze the content from image, text, audios. It helps to generate medical data to training the clinical trials and new drug characteristics, its side effects and usages. The medical data such as drug names on sheets can be parsed as input image, therefore extracts text from the image based using tesseract a highly enhanced open source. After extracting the image transmute it into a text file, which will be source of truth for the Generating summary and analytical theory from the text, which therefore gives side effects, usages and similar drug lists. This will be greatly applicable for the persons who have limited knowledgein understanding drugs, they can simply get end to end results.

2. RELATED WORK

In current research on healthcare industry, we use to retrieve user health data based on his activity [1] on Internet of things. These data can be used for further analysis to design a model upon his activity, with CNN model. Artificial intelligence supports in understanding and create a support system with use of deep learning techniques. Common methodologies in analytical representations of health care data are logistic regression, support vector machine and random forest.
In [1], the authors proposed a CNN model to identify the patterns in understanding in the user health status as it is support to life care based on living pattern analysis. In [10] authors proposed a method to do pattern analysis on health care which constantly changes based on the activity of the user. These challenges can be solved by various ways or techniques using machine learning models, based on predictive analysis.Earlier [3] hospitals used to treat the patients by writing prescriptions on paper and by words. Now a day's thing changed to great extent where everything can be saved as digital information where it can be used to diagnosis the person diseases. Based on EMR data machine learning algorithms help in diagnosing further, the doctor and patient order information can be survived as input for treatment pattern discovery. In the [2] research work, authors classified a novel approach to determine heart disease identification methods using classification algorithms such as Logistic regression, support vector machine, Artificial Neural Networks, K- nearest neighbor, Naïve bays, and Decision tree. Among these algorithms, data is more efficiently performed in a novel fast conditioning algorithm in identification of heart diseases which is one of the complex diseases to identify. The detailed experimental results show feature-based selection algorithms. These suggested algorithms have obtained a great accuracy as compared to earlier methods in classification. Rehan Ashraf, Muhammad Asif Habib et al (2020) [5] discussed to use deep learning models in describing the health care related records based on big data medical image classifications for detection. Eslam Amer,Abdelmgeid A. Ali et al (2021) [6] to classify the medical data using K- nearest neighbor in classification of medical tasks. Ardhendu Sekhar,Soumen Biswas et al [7] together proposed a fine-tuned GoogleNet algorithm based in identifying the early signs of brain tumor in cascading the dataset to refine the algorithm. Mengfang Li,Yuanyuan Jiang et al (2023) [8] discussed
through an systematic organisational updates for enhancing the model, these leaves a greater study in across of all machine learning theory. Binila Mariyam

Module	Purpose
Tesseract	To extract the text from the image.
Pre-processing	Input the dataset to train the algorithm for generativeAI.
Working principle,	Flow of model,how to handle the dataset and train the algorithm.
Generative AI	GenAI to get the information be generated from the Drug name discovery.
Data augmentation	Re augmenting the data set for training the models with larger volume of data to improve test data accuracy.

Boban,Rajesh et al (2020) [9] proposed the model which improves accuracy in handling lung diseases classification based on machine learning algorithms.
Appoorva Bansal,Anand kr.Shula et al (2022) [10] machine learning provides a greater visibility to produce the model with a huge dataset. C. Han, L. Rundo, K. Murao et al (2020) [11] research provide a comprehensive solution for all machine learning algorithms to enhance which can be used to repurpose and extend the model to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

greater products in healthcare. Most Literature studies states that generative AI resolves and been a greater mode of Upgradation for current systems. Y.-J. Cao, L.-L. Jia (2008) [12] These makes product quality to be improved and states a outcomes in enhancing the possibilities in new world. Y.-H. Nho,
S. Ryu (2021) [13] Artificial intelligence and machine learning immerses the healthcare domain to emerging trends in improving the possibilities for making the world better place.

G. Lan, S. Xiao et al (2023) [16] proposed a data driven Artificial Intelligence model to improve the cognitive features in enhancing the ability to identify and all the abnormal outliers identified to perform the fine-grained enhancement. The following summaries all the quests will be fine tuned using GenAI - Data driven approach. K. Moulaei, A. Yadegari et al (2024) [17] solutions for complex healthcare problems to resolve based on Generative AI which integrates all the factors in healthcare domain for reassessing the applications, benefits and challenges in healthcare. These literatures provided a high insights in handling the complex problems to find the better solution in healthcare. As a result we proposed an model in handling the images of drug information to perform better analytical approach. Our model summarises in taking the technical considerations as part of current ongoing research and emerging trends in healthcare. We implemented a solution for enhancing the drug images in real time, extract the information present behind the drug. The extracted information is further calculated and captured into a txt file to understand and train our model with machine learning methods. Among all the machine learning algorithms that were captured in our further study in coming sections, support vector machine is most accurate algorithm in providing greater solutions for the images to text converted snippets. The generative AI feature calibrated further to get the usage of the drugs and the side effects of using from the trusted sources. Retrieving this information and providing a summary can help the end user to find the better drug by himself on summarizing the information. In regard to [21][22], the storage used to be optimal and cheaper rates are considered for storing the insights, in that way, these studies would benefit in choosing practices for optimization over the cloud.

3. Methodology

The modules identified are Tesseract, Pre-processing, Working principle, Generative AI , and Data augmentation.

Table 1. Modules purposes

3.1 Tesseract:

Tesseract is a famous extraction mechanism where it extracts the text from the images, and documents, and converts them into text format files in our case.

This is a popular engine for Natural language processing, which uses only highly resolution images for processing. It recognizes various languages; Tesseract can be used as a diverse multilingual application. Handwritten languages also can be recognized but with a low accuracy compared to highly resolution images.
This is because of different handwriting styles by different persons.

3.2 Data Download and Pre-processing:

The pre-processing images were processed during run time and the model has been trained with pretext classifications with the labelling of data. Every image processed is in the run time, this helps not to consume more space or storage for retrieving the output.

3.2.1 Working Principle:

Medical data needed to be scaled and classified correctly, so that no one gets confused in classifying them. The data obtained after thoroughly pre-processing with tesseract and machine learning algorithms.

PS1: Pseudo_procedure
Classification_report(Images[][], text[][]):
Input: Image
Output: Text report

Step 1: Input a highly resolution image as input for better results.
Step 2: Adaptive thresholding for extracting text from image.
Step 3: Results finding all the characters in the image and save in .txt format.
Step 4: Preprocess it with logistic regression and Support Vector Machine Algorithms to classify the medical and non- medical data.
Step 5: Computing regression coefficients for the training data of medical and non-medical data classification.
Step 6: Finding the relationship between training and testing data, using the SVM and Logistic Regression algorithms.

P(Y=1|X) =1/ 1+e−(β0+β1X1+β2X2+⋯+βnXn)

Step 7: The results declared after rigors classification between the algorithms to

$$w^T(\Phi(x)) + b > 0$$

identify the best accurate algorithm.
Step 8: Iterate from step 2 to step 7 for every image and train the model.

Logistic Regression is the way statistically where it predicts binary variable. This gives binary output based on the below mathematical function where it maps directly predict the values to probabilities:
P(Y=1|X) =1 / 1+e−(β0+β1X1+β2X2+⋯+βnXn)

Where Y is the dependent variable, X represents the independent variables and β it is the coefficients.

Understanding support vector machine algorithm provides higher accurate results in classifying the medical health care data and non-medical health care data upon converting the image into text, preprocessing the data over each coefficient of training data in classification and regression. It can easily produce distinguish between two groups and provide the probability of values with accuracy, below is the algorithm used for providing SVM.

Here we have to process the function in identifying the positive group. The predictions if any group is considered as negative it follows with the prediction of this and make those particular values as negative value group.

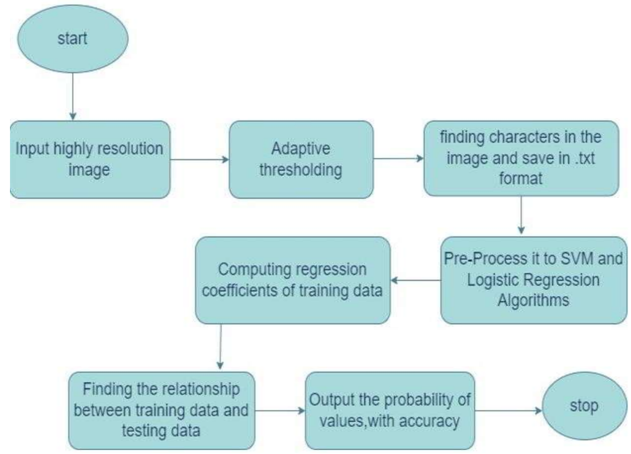

Fig.1. High Level Data Approach

Flow needed to be visualized and overviewed deeply so any beginner starting with knowing all the research methods can be known. Classifying medical and non-medical data we need to start inputting a high-resolution image to the model. This AI model will use Tesseract an optical character recognition tool to adapt threshold images for finding characters in the image and capture highly accurate results by saving in .txt format. Once the data has been saved, it will be used as input for the next part of the algorithmic approach. Pre-processing the input with SVM and Logistic Regression Algorithms, these algorithms provided with computing of regression coefficients of training data. Using both the algorithm relationship between training and testing data is identified and captured in the next Result sections. The output includes all the probability values, along with the accuracy.
The following are the uses after the proposed system is applied to the scanned medical image extraction.

1. Personalized Medication Plans: Generative AI can be used for performing statistics using current trends, and lifestyle factors, to create tailored medication plans. This can help ensure that patients receive the right dosage at the right time.

Example: A generative model could take input

$$w^T(\Phi(x)) + b < 0.$$

data about a patient's health conditions and generate a customized medication schedule that includes dosage, frequency, and any necessary

Journal of Theoretical and Applied Information Technology
31st January 2025. Vol.103. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

adjustments based on the patient's response.

2. Medication Reminders: AI can generate reminders for patients to take their medications. These reminders can be personalized based on the patient's routine and preferences, increasing adherence to medication schedules. Example: A chatbot or mobile app could use generative AI to create daily reminders that are sent to the patient via text or app notifications, including motivational messages or tips on managing side effects.

3. Educational Content: Generative AI can create informative content about medications, including potential side effects, interactions, and proper usage. This can help patients understand their treatment better and encourage adherence. Example: An AI model could generate personalized educational materials, such as pamphlets or videos, that explain how to take a specific medication, what to expect, and when to seek help.

4. Simulation of Medication Effects: Generative models can simulate how different medications might affect a patient based on their unique profile. This can help healthcare providers make informed decisions about prescribing. Example: A generative model could analyze historical data to predict how a new medication might interact with a patient's existing medications and conditions, providing insights into potential outcomes.

5. Feedback and Adjustment: Generative AI can also facilitate ongoing feedback from patients about their medication experience, allowing for real-time adjustments to their medication plans.

Example: Patients could provide feedback through an app, and the AI could generate suggestions for adjustments based on their reported side effects or the effectiveness of the medication.

6. Implementation Considerations:

Data Privacy: Ensure that patient data is handled securely and in compliance with regulations like HIPAA. User-Friendly Interfaces: Develop intuitive interfaces for patients to interact with the AI system.

Collaboration with Healthcare Professionals: Work closely with healthcare providers to validate the AI-generated plans and content.

3.3 Generative AI: Soon after retrieving the output.txt from the image, identifies the drug name and searches in real time to get the summarized information from Government approved websites. This uses the Google generative AI keys to obtain the API information and hit the respective server. Due to which it gets report from the real time updated websites, so that we can ensure our data obtained is always latest. Using Google APIs in the code gives the real time information unless it reduces the training of our algorithm. Which can reduce the cost and effort in training the data. Only when classification of the input image is completed then fetch the .txt file and source that input to Gen AI algorithm. This provides the resulted summarized report.

3.4 Data Augmentation: Due to the fact that our model requires new generative data, it requires training the model with existing data and produce new information. This results better resizing the dataset and balance the class information with different classifications or transformations. This precisely results better architecture of any model, when implementing the Data augmentation model. It will improve and generalize the dataset for better results. Producing the drug discovery in any means of way can be helpful for any person, because before taking any medicine, one need to understand what is the use of the drug. Every person should have an understanding why, how, when for any drug they are recovering with. This kind of similar model has been implemented with a clear way of examples in coming sections.

4. EXPERIMENTAL RESULTS

In this, investigation phase is initated, then classification report using logistic regression, then generation of report using SVM, and then generation of report using Generative AI.

A. Investigations: To extract text from an image we used Tesseract OCR and PyTesseract. The Fig 2 tells about at runtime. It will asks us to choose file or else we can cancel upload .so it will prompt the user to upload image file using 'files.upload' function.

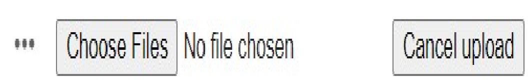

Fig.2. Input Image

Fig.2 and Fig.2.1 is the input image where our model will start analyzing. We have uploaded animage because we need to extract a text from

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

image.

Fig.2.1. Uploading the image

```
Choose Files  WhatsApp I....06 PM.jpeg
• WhatsApp Image 2024-04-02 at 2.52.06 PM.jpeg(image/jpeg) - 117534 bytes, last modified: 4/3/2024 - 100% done
Saving WhatsApp Image 2024-04-02 at 2.52.06 PM.jpeg to WhatsApp Image 2024-04-02 at 2.52.06 PM.jpeg
```

Fig.3 is the sample input image of our experiment where it is a backside image of medicine. Herewe can upload any image based on user choose. Make sure we upload only image.

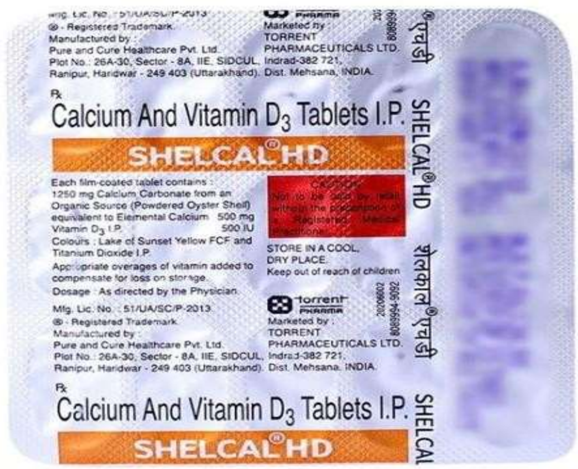

Fig.3. Backside image of medicine

```
output.txt
1 Mig, Ue-No.:StIVA'SCIP-2013
2 © -Repisteroa Taderark
3 Manufactured by:
4
5 omy
6 Pot No. 264.30, Setor- 2 IE, SIOCUL, Indra982 721,
7 FRanpur, Mardwar 249 403 (Utarathand), Dist Mehsana,
8
9 Calcium And Vitamin D Tablets LP.
10
11 SHELCACHD
12
13
14
15 WOT:
16 FF
```

Fig.4. Output extracted from image

From Fig.4 we extract text from the image by using PyTesseract. Save the extracted text to output.txt. To print the stored output. We used 'stored_output'.

```
Predicted Label: medical
```

Fig.5. Prediction Labels

We have preprocessed to remove any irrelevant information and then converted into numerical vectors using techniques such as TF-IDF or Bag of Words. Next these numerical vectors are fed into the Logistic Regression and SVM algorithms to train our model. Here Fig.5 the Logistic Regression algorithm uses sigmoid functions to predict the probability of an output. Text files belonging to the medical or non-medical class. Here we uploaded image of medicine so we got output as medical.

```
Predicted Label: medical
Accuracy: 0.5
Precision: 0.25
Recall: 0.5
F1 Score: 0.3333333333333333
```

Fig.6. Consolidated results using Logistic Regression

From Fig.6 we have evaluated the performances of the model by using logistic Regression. We have used these algorithms to estimate the effectiveness, and precision in the training, recall, and overall F1-score. Accuracy is the percentage of accurately classified text documents. Precision is the % of true binary positive values among total predictions of positive cases. Recall is the percentage of true positives among all actual positives. The other measure F1-score is also computed as precision and recall average.

Classification Report:

	precision	recall	f1-score	support
medical	0.67	1.00	0.80	2
non-medical	1.00	0.50	0.67	2
accuracy			0.75	4
macro avg	0.83	0.75	0.73	4
weighted avg	0.83	0.75	0.73	4

Fig.7. Classification Report

From Fig.7 we can understand the predicted label for Logistic Regression is listed above with the classification report of precision, recall, f1-score, and accuracy. The overall accuracy for the logistic regression algorithm is

Journal of Theoretical and Applied Information Technology
31st January 2025. Vol.103. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

0.5.

```
Predicted Label: ['medical' 'medical' 'medical' 'medical' 'medical' 'medical' 'medical'
 'medical' 'medical' 'medical' 'medical']
SVM Evaluation Metrics:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
```

Fig.8. Classification report for SVM

From Fig.8, we also used SVM algorithms to train the model. The SVM algorithm, on the other hand, builds a high dimensional space, and defines a hyperplane which determines output into two classes. The SVM algorithm achieved an accuracy of 1.0, precision, recall, and F1-score of 1.0.

Line by line predicting labels of medical and non-medical data, overall, the input image is medical data; hence the object is predicted as medical with the metrics. Based on the results, it appears that the SVM algorithm outperforms the Logistic Regression algorithm for medical and non-medical text classification in terms of accuracy, precision, and F1- score.

Going too deep in further analysis, we propose the hybrid modelto include the Generative AI solution to look up in Google and bring back the information for any drug. This hybrid model summarizes all the information which in return helps to get more data, to emphasize the anlysis.

```
Disclaimer: This is for informational purposes only. Consult a doctor for accurate information about Dolo 650.
**Search result from Paracetamol usages regulation: A need of the hour - PMC**
Nov 1, 2022 ... This medication has recently gained attention in India. Dolo-650 tablet, a commonly used medicine among Indian households
More information: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9663334/
```

Fig.9. Gen AI output for Dolo-650 drug

Fig.9 retrieve the summary for Dolo-650 drug, which is accurate to its usage and it also provided an more information url to look for the extended information on govt approved websites. So, this leaves an image behind that it can be trustworthy in obtaining the information, if required we can browse to that url providedin the end.

```
Epoch 38/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 22ms/step - kl_loss: 1.9343 - loss: 357.0987 - reconstruction_loss: 355.1643 - val_loss: 0.0326
Epoch 39/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 22ms/step - kl_loss: 1.8504 - loss: 354.0504 - reconstruction_loss: 352.2000 - val_loss: 0.0329
Epoch 40/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 22ms/step - kl_loss: 1.9253 - loss: 359.3054 - reconstruction_loss: 357.3800 - val_loss: 0.0323
Epoch 41/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 28ms/step - kl_loss: 1.9192 - loss: 354.2981 - reconstruction_loss: 352.3790 - val_loss: 0.0325
Epoch 42/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step - kl_loss: 1.9199 - loss: 347.8053 - reconstruction_loss: 345.8854 - val_loss: 0.0310
Epoch 43/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 23ms/step - kl_loss: 1.9866 - loss: 351.9084 - reconstruction_loss: 349.9218 - val_loss: 0.0307
Epoch 44/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 23ms/step - kl_loss: 1.9093 - loss: 350.6846 - reconstruction_loss: 348.7754 - val_loss: 0.0309
Epoch 45/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step - kl_loss: 1.9432 - loss: 347.6205 - reconstruction_loss: 345.6773 - val_loss: 0.0303
Epoch 46/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 28ms/step - kl_loss: 1.9686 - loss: 346.2946 - reconstruction_loss: 344.3260 - val_loss: 0.0306
Epoch 47/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 22ms/step - kl_loss: 1.9541 - loss: 345.4798 - reconstruction_loss: 343.5257 - val_loss: 0.0302
Epoch 48/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step - kl_loss: 1.9776 - loss: 339.7908 - reconstruction_loss: 337.8132 - val_loss: 0.0297
Epoch 49/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step - kl_loss: 1.9452 - loss: 342.7415 - reconstruction_loss: 340.7963 - val_loss: 0.0297
Epoch 50/50
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step - kl_loss: 1.9652 - loss: 341.7673 - reconstruction_loss: 339.8021 - val_loss: 0.0298
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
```

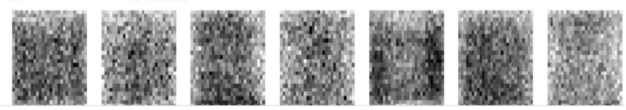

Fig. 10. Epochs and correlation between run time

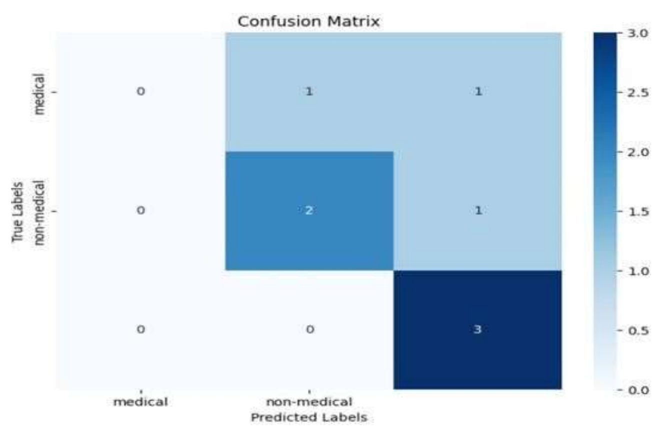

Fig.11. Confusion matrix for medical and non-medical

The confusion matrix shown in Fig.11 talks about the model correctly classified a certain number of texts as 'medical', 'non-medical' and 'borderline' as shown by diagonal values.

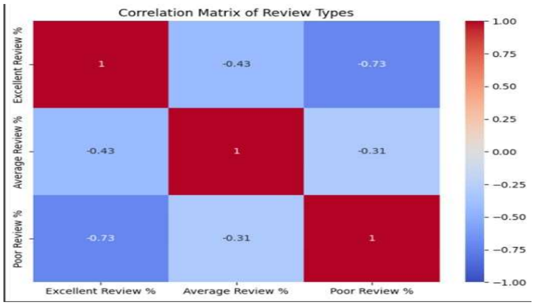

Fig.12. Correlation Matrix of review types of medicine

From Fig.12, illustrates the relationships between three different types: 'Excellent Review%', 'Average Review %', 'Poor Review'. Each cell in the matrix shows the correlation between two review types, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

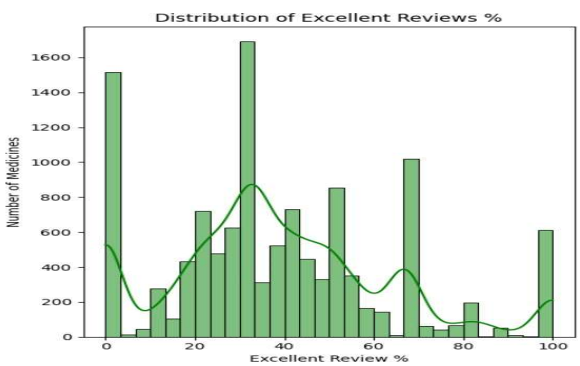

Fig.13. Histogram for distribution of excellent reviews of medicines

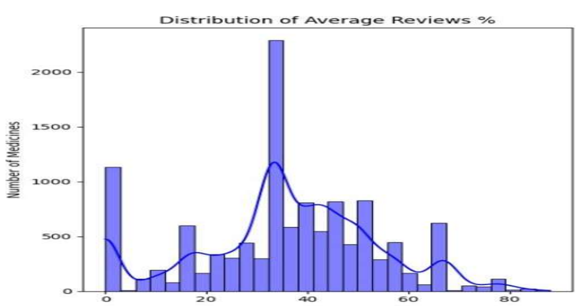

Fig.14. Histogram for distribution of Average reviews of medicines

From Fig 13,14,15 based, histograms were drawn from datasets to understand how the medicines are perceived in terms of quality, as indicated by percentages of excellent and poor reviews. These provide insights into how medicines are rated by users. Where from Fig.12 we got know that distribution with multiple peaks, indicating that there is variability in how medicines are rated as excellent.

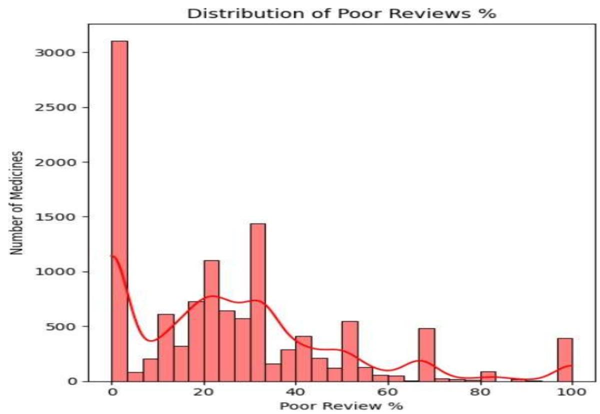

Fig.15. Histogram for distribution of poor reviews of medicines.

Some medicines got very high percentages of excellent reviews while other medicines are clustered around low percentages. The KDE curve tells that distribution is not uniform and has several distinct groups, that different subsets of medicines are perceived differently in terms of excellence.

From Fig 15 gives 'poor reviews %' is skewed to left, meaning that the most of medicines receive a low percentage of poor reviews. Some significant concentration of medicines with 0% to 20% poor reviews, indicating the most medicines are not frequently rated poorly.

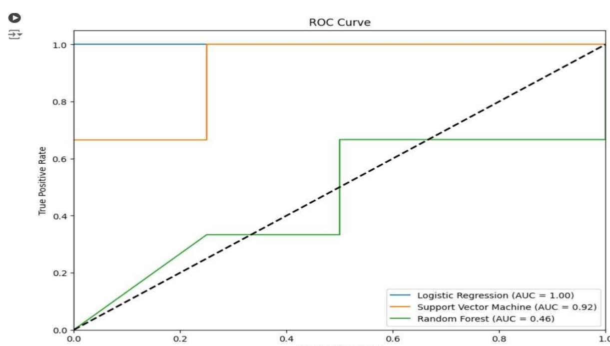

Fig.16. ROC curve

From Fig.16 we compare logistic regression (AUC=1.00), SVM(AUC=0.92) and Random Forest (AUC=0.46). Logistic regression shows prefect classification, SVM performs well, while Random Forest is nearly random. The closer the curve is to the top-left, the better the model.

Table 1 demonstrates evaluation of specific measures accuracy, precision, recall, and F1-score of hybrid model(Logistic Regression,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Support Vector machine, and Data Augmentation) against Naïve Bayes, Decision, Trees, Generative AI.

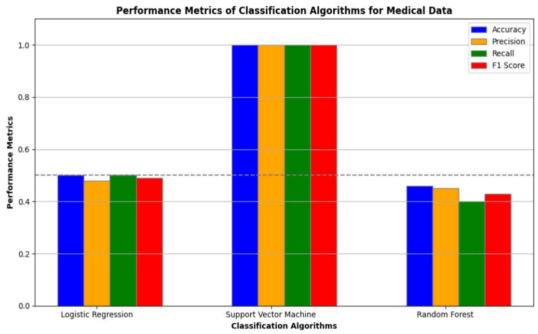

Fig.17. Performance metrics of classification algorithms for medical data

The visual demonstration of Table 1 is depicted in Fig.18 that shows comparison of accuracy, and performance.

Table 1. Measures against considered methods

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Naïve Bayes	82	79	74	76.5
Decision Trees	85	81	78	79.5
Random Forest	88	83	80	81.5
Generative AI	97	94	92	93.0
Hybrid Method (Logistic Regression + SVM + Data Augmentation)	100	99	95	96.5

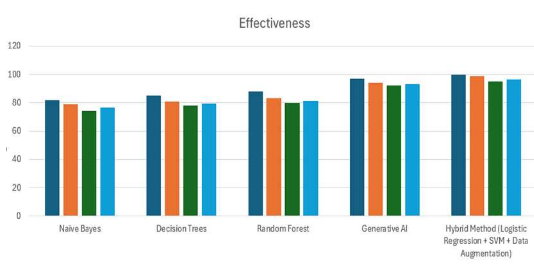

Fig.18. Effective of considered models on classification

Overall performance metrics for all algorithms used in our model, is captured to understand better visually. Support Vector machine algorithm have given significant results. Furthermore, identifying the drug name from the image and generating the report on its uses and disadvantages will support users without worrying of reading and searching separately over internet.

5. CONCLUSION AND FUTURE SCOPE

Based on our research we are classifying the logistic regression and support vector machine algorithms for medical and non-medical data classification, also creating a summarized information regarding any kind of drug from the govt approved websites. The results and analyzing showed that support vector machine performed better in terms ofaccuracy.

This suggests that support vector machine is a more suitable algorithmic approach for this classification task. The results indicate that support vector machine can provide more accurate identification and categorization of medical and non-medical data. Therefore, it is recommended to use support vector machine for this specific classification problem. Overall, the use of machine learning algorithms for text classification can help automate and improve the accuracy of identifying and categorizing medical and non-medical data. In the end, for obtaining the information we have to use Generative AI, which gives highly valuable information or report from the encyclopedia. One can extend this model in day-to-day tasks for workflows and the classification of algorithms gives better results, which will miss in identifying in human sight. The generative AI can be extended to include this model in Application feature to retrieve the analytical information.

REFERENCES

[1] Ismail, W. N., Hassan, M. M., Alsalamah, H. A., & Fortino, G. (2021), CNN-based health model for regularhealth factors analysis in internet-of-medical things environment. IEEE Access, 9, 157108- 157118.

[2] Jian Ping Li,Amin Ul Haq,Salah Ud Din,Jalaluddin (2020), Heart Disease Identification Method Using Machine Learning Classification Healthcare, IEEE10.1109/ACCESS.2020.3001149.

[3] Jahanzaib Latif,Chuangbai Xiao,Shanshan Tu,Sadaqat Ur Rehman,Azhar Imran,Anas Bilal(2020), Implementation and Use of Disease Diagnosis Systems for Electronic Medical Records Based on Machine Learning: A Complete Review.IEEE,10.1109/ACCESS.2020. 3016782.

[4] Andreas S. Panayides,Amir Amini,Nenad D. Filipovic,Ashish Sharma,Sotirios A. Tsaftaris,Alistair Young,David Foran,Nhan Do,Spyretta Golemati,Tahsin Kurc,Kun Huang,Konstantina S. Nikita,Ben P. Veasey,Michalis Zervakis,Joel H. Saltz,Constantinos S. Pattichis(2020), AI in Medical Imaging Informatics: Current Challenges and Future Directions, IEEE, 10.1109/JBHI.2020.2991043

[5] Rehan Ashraf,Muhammad Asif Habib,Zahoor Abbas,Muhammad Yasir,Toqeer Mahmood,Saadat Hanif Dar,Muhammad Awais,Muhammad Sheraz Arshad Malik(1 june 2020), Deep Convolution Neural Network for Big Data Medical Image Classification, IEEE, 10.1109/ACCESS.2020.2998808

[6] Eslam Amer,Abdelmgeid A. Ali,Essam H. Houssein,Eman Saber,Yaser M. Wazery (2021), An Efficient Slime Mould Algorithm Combined With K-Nearest Neighbor for Medical Classification, Tasks,IEEE, 10.1109/ACCESS.2021.3105485.

[7] Ardhendu Sekhar,Soumen Biswas,Ranjay Hazra,Arun Kumar Sunaniya,Amrit Mukherjee,Lixia Yang (2021), Brain Tumor Classification Using Fine-Tuned GoogLeNet Features and Machine Learning Algorithms: IoMT Enabled CAD Systems, IEEE, 10.1109/JBHI.2021.3100758

[8] Mengfang Li,Yuanyuan Jiang,Yanzhou Zhang,Haisheng Zhu(2023), Medical image analysis using deep learning algorithms, https://doi.org/10.3389/fpubh.2023.1273253.

[9] Binila Mariyam Boban,Rajesh Kannan Megalingam(2020), Lung Diseases Classification based on Machine Learning Algorithms and Performance Evaluation, IEEE ,10.1109/ICCSP48568.2020.9182324

[10] Appoorva Bansal,Anand kr.Shula ,Shreya Bansal(2022) Machine Learning Methods for Predictive Analytics in Health Care, IEEE,10.1109/SMART52563.2021.9676233.

[11] C. Han, L. Rundo, K. Murao, T. Nemoto, and H. Nakayama, "Bridging the Gap Between AI and Healthcare Sides: Towards Developing Clinically Relevant AI-Powered Diagnosis Systems," in Artificial Intelligence Applications and Innovations, vol. 584, I. Maglogiannis, L. Iliadis, and E.P imenidis , Eds . Cham ,Switzer land :Springer ,2020,pp .320 - 333 , doi : 10.1007/978-3-030-49186-4_27.

[12] Y.-J. Cao, L.-L. Jia, Y.-X. Chen, N. Lin, C. Yang, B. Zhang, Z. Liu, X.-X. Li, and H.-H. Dai, "Recent Advances of Generative Adversarial Networks in Computer Vision," IEEE Access, vol. 7, pp. 14985- 15006, Dec. 2018, doi: 10.1109/ACCESS.2018.2886814.

[13] Y.-H. Nho, S. Ryu, and D.-S. Kwon, "UI-GAN: Generative Adversarial Network-Based Anomaly Detection Using User Initial Information for Wearable Devices," in IEEE Sensors Journal, vol. 21, no. 8, pp. 9949-9958, Apr. 2021, doi: 10.1109/JSEN.2021.3054394

[14] C. T. Nguyen, Y. Liu, H. Du, D. T. Hoang, D. Niyato, D. N. Nguyen, and S. Mao, "Generative AI- enabled Blockchain Networks: Fundamentals, Applications, and Case Study," in IEEE Network, Early Access, pp. 1-1, Jun. 2024, doi: 10.1109/MNET.2024.3412161.

[15] A. Khaled, J.-J. Han, and T. A. Ghaleb, "Multi-Model Medical Image Segmentation Using Multi- Stage Generative Adversarial Networks," in IEEE Access, vol. 10, pp. 28590-28599, Mar. 2022, doi: 10.1109/ACCESS.2022.3158342.

[16] G. Lan, S. Xiao, J. Yang, J. Wen, and M. Xi, "Generative AI-based Data Completeness Augmentation Algorithm for Data-driven Smart Healthcare," in IEEE Journal of Biomedical and Health Informatics, Early Access, pp. 1- 8, Oct. 2023, doi: 10.1109/JBHI.2023.3327485.

[17] K. Moulaei, A. Yadegari, M. Baharestani, S. Farzanbakhsh, B. Sabet, and M. R. Afrash, "Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications," in International Journal of Medical Informatics, vol. 188, pp. 105474, Aug. 2024, doi: 10.1016/ j.ijmedinf.2024.105474.

[18] S. Sai, A. Gaur, R. Sai, V. Chamola, M. Guizani, and J. J. P. C. Rodrigues, "Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations," in IEEE Access, vol. 12, pp. 31078-31106, Feb. 2024, doi: 10.1109/ ACCESS.2024.3367715.

[19] P. Zhang and M. N. Kamel Boulos, "Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges," Future Internet, vol. 15, no. 9, p. 286, Aug. 2023. doi: 10.3390/

fi15090286.

[20] H. Ali, M. H. Rehmani, and Z. Shah, Eds., Advances in Deep Generative Models for Medical Artificial Intelligence. Springer Cham, 2023, pp. XVI, 248. doi: 10.1007/978-3-031-46341-9.

[21] H.K.R. Dey, N.S., Sangaraju(2024), A particle swarm optimization inspired global and local stability driven predictive load balancing strategy,Indonesian Journal of Electrical Engineering and Computer Science, 35(3), 1688-1701, DOI: 10.11591/ijeecs.v35.i3.

[22] NS Dey, HKR Sangaraju(2023),Hybrid Load Balancing Strategy for Cloud Data Centers with Novel Performance Evaluation Strategy, International Journal of Intelligent Systems and Applications in Engineering, 11(3), https://ijisae.org/index.php/IJISAE/article/view/3345.