Impact Factor (SJIF): 5.158 E-ISSN: 2583-7206

International Journal of All Subject Research

Received: 22/December/2024 IJASR: 2025; 4(1):04-10 Accepted: 31/January/2025

Interactive Exploratory Data Analysis (EDA) Dashboard: Advanced Security, Auto-ML, and Data Storytelling

[*1]D. Aashish, [2]Ch. Sai Sri Saran and [3]N. Senthamilarasi

[*1, 2]Student, Department of Computer Science and Engineering with Artificial Intelligence and Machine Learning, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India.

[3]Assistant Professor, Department of Computer Science and Engineering with Artificial Intelligence and Machine Learning, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India.

Abstract

This paper introduces us to a new age technological advancement in the field of healthcare with AI by developing an embedded system comprising of healthcare tools with a dashboard that works based on various modules embedded for its effective working like GPT-based transformers, Langchain, scikit-learn for Auto-ML, Exploratory Data Analysis, NLP for identifying the medical parameters, OCR technology etc. The healthcare tools contain the organ based disease diagnosing system that diagnoses the condition of the impact region and predicts the possible condition of the organ, Medical PDF reports analyser, CSV file analyser and a healthcare chatbot for regular common questions posed by the patients and a dashboard that provides us a comparative visual representation of the actual level of various healthcare parameters like (Blood pressure, allergy levels etc.) in patients versus its normal levels where it should exist. The dashboard presents data through engaging visualizations such as scatter plots and donut charts. The dashboard also contains interactive widgets that are in the colour according to the deviation in the level of parameters or KPI's in the patient. Its colour coded based on four colours red, amber, yellow and green. The dashboard also provides us recommendation for workouts, dietary changes, precautions to be taken, medications that can be taken and how to bring the levels under control based on each parameter level analysis. Keywords-Auto-ML, GPT-Based Transformers, OCR Technology, scikit-learn, Exploratory Data Analysis, NLP, healthcare tools, Dashboard.

Keywords: EDA Dashboard, Advanced Security, Auto-ML, Technological Advancement etc.

1. Introduction

This paper is centered on the creation of healthcare tools powered by AI to analyse the pdf format medical lab reports, csv based medical reports, question-answer bot for answering patients queries regarding diseases, precautions, medications etc., Xray scan analysis system that uses image processing tool to find the impact region in the organ and diagnose the possible problem by using predictive features like AutoML, Exploratory Data Analysis, NLP, Computer vision etc. Streamlit tool used is an open source library used in python to provides highly interactive User interface for the web application developed. Its key feature is the simplified and a compact code design. We use a variety of OpenAI models and transformers in our models including GPT-3.5 turbo, GPT-4 and DALL-E (Image generation model). Natural Language Processing is used to perform Named Entity Recognition to differentiate between Health parameters and Non health parameters. We use deep learning based models like ResNet, ImageNet, VGGNet, Convolution Neural Networks with multiple layers of combination of convolution and pooling layers to derive deeper features of the image and classify with a higher accuracy, Recurrent Neural Networks are also used for prediction purpose.

The Interactive dashboard that is being developed is comprising of streamlit, interactive widgets, Plotly to provide attractive visualizations, Exploratory data Analysis for the data that involves the data preprocessing steps of data collection, data cleaning, data validation, data analysis, model evaluation, model visualization, model deployment, types of analysis done include Time series analysis, Correlation analysis, Outlier detection for the healthcare data to get deeper insights about the symptoms and try to compare between the real time conditions similarities between different patients to give effective treatment to the patient. The dashboard provides an additional data storytelling component that transforming raw data into narrative by identifying key patterns in the data, analyzing data, tailoring content, creating visually attractive charts and graphs that provide us a 360 degree insights about the patients condition enabling the doctors to understand the situation carefully. The dashboard also gives us an feature of displaying the value of healthcare parameters or KPI's (like Allergy, Blood Pressure) in colour

***Corresponding Author:** D. Aashish

coded widgets where each colour signify the risk level. It also provides clinical notes about the diseases the patient is in the verge of risk and to act cautiously. The red colour signify the patient is seeing highly hazardous levels of a parameter, Amber means moderate risk category, Yellow means low risk, Green means no risk.

2. Literature Survey

AutoML capabilities can be incorporated into dashboards to improve result accuracy by automating model selection and optimization processes. This enhances dashboard functionality, making it easier to leverage automated processes for predictive analytics and decision support [1].

Data visualization literacy is crucial for the effective use of reporting dashboards. Users often face challenges in interpreting visual data. Improving literacy involves using intuitive layouts, clear labeling, and context-aware visualizations to bridge the gap between data complexity and user comprehension, thereby enhancing decision-making [2].

Combining visual components with concise narratives helps summarize data effectively. This methodology focuses on key insights, making complex data more accessible and understandable by presenting it in a clear and structured format [3].

Security Monitoring as a Service (SMaaS) is a security solution tailored for cloud-based analytic applications. It addresses unique challenges like dynamic threat landscapes and shared infrastructure risks through real-time monitoring, anomaly detection, and adaptive security mechanisms, providing a scalable and efficient solution for enhancing cloud security [4].

AutoML models are designed to process large-scale industrial data by automating data explanation and analysis. This reduces manual intervention, enabling organizations to derive actionable insights from complex datasets efficiently [5].

Developing dashboards for organizational analytics involves addressing security challenges. Essential measures include implementing robust security protocols, regular bug assessments to prevent data leakage, and incorporating advanced data security tools. Data encryption methods also play a crucial role in safeguarding sensitive organizational information [6].

Effective data preprocessing techniques are essential for maximizing the potential of AutoML models. This involves cleaning, transforming, and normalizing data to develop accurate and reliable machine learning models, enhancing their performance [7].

Data Scope is a visual analytics dashboard designed to handle large and multidimensional datasets. It focuses on creating attractive visualizations that aid in analyzing complex relationships between dataset attributes. The integration of interactive visual tools provides a seamless user experience while managing scalability and high-dimensional data challenges [8].

Incorporating security measures into data visualization processes is crucial. This involves integrating real-time threat detection systems within visualization tools, ensuring the integrity and confidentiality of analyzed data [9].

Tools like Tableau and Plotly enable users to derive critical insights by presenting data in an intuitive and visually appealing manner. User-centric design is essential for facilitating data interpretation and decision-making [10].

AutoML tools offer advantages such as increased efficiency and accessibility, but they also come with challenges like computational resource demands and limitations in handling highly complex datasets. Optimizing AutoML implementations ensures a balance between automation and user control [11].

Exploratory Data Analysis (EDA) is critical for providing valuable insights for professionals involved in data-driven engineering and software development tasks. Visual narratives simplify complex data structures and facilitate decision-making processes. EDA helps identify trends, outliers, and patterns within datasets, empowering engineers and developers to design effective solutions based on data-driven insights [12].

Integrating big data analytics into cybersecurity systems enhances security features by analyzing vast amounts of data for patterns and anomalies indicative of cyber threats. This approach emphasizes the role of big data-driven cybersecurity systems in proactive threat detection and mitigation [13].

A Vulnerability Management Dashboard is designed to identify and mitigate security risks during data analysis. It provides security analysts with tools to assess vulnerabilities and make informed decisions regarding system enhancements. Features include automated vulnerability detection, risk scoring, and actionable recommendations for improving system security [14].

AutoML addresses challenges in human oversight during model development by optimizing model selection based on metrics such as error percentage, accuracy, and loss rates. By automating these processes, AutoML makes machine learning tools accessible to users without programming expertise, maximizing tool efficiency and reducing human-induced errors [15].

Effective data storytelling involves using attractive visualizations, minimal textual content, and storytelling elements like images and diagrams to engage audiences. Structuring data stories maximizes their impact on diverse audiences [16].

Practical techniques for data storytelling in big data analytics focus on sharing insights effectively. Creating relatable narratives and leveraging visual storytelling tools emphasize the importance of storytelling in bridging the gap between raw data and actionable insights [17].

DynDash is a multi-coordinated dashboard designed to perform data visualizations from various perspectives. The coordination between different visualization components facilitates users in gaining a holistic understanding of datasets. This approach is particularly effective for analyzing datasets with intricate interdependencies, making DynDash a powerful tool for data analysts and researchers [18].

AI-powered diagnostic systems have emerged as a transformative solution in healthcare, leveraging GPT-based transformers and Auto-ML frameworks for enhanced disease detection. These systems address critical challenges like processing complex medical imaging and large datasets through advanced algorithms, ensuring precise diagnosis and early intervention. By integrating real-time analytics and predictive models, they offer scalable and efficient mechanisms to detect organ-specific abnormalities and provide tailored treatment recommendations, revolutionizing precision medicine [19].

Advanced healthcare dashboards serve as a comprehensive solution for patient monitoring and data interpretation, utilizing EDA and interactive visualizations to address challenges such as real-time tracking of health parameters and anomaly detection. By employing tools like scatter plots and color-coded widgets, these dashboards simplify the understanding of deviations in key health metrics. Integrated

with IoT devices and predictive analytics, they provide actionable insights, including dietary suggestions, medication adjustments, and lifestyle recommendations, ensuring a proactive and personalized approach to healthcare management [20].

3. Process and Tools

The project focuses on building a tool for healthcare sector by combining various new age technologies like Artificial Intelligence AI, Machine Learning ML, Neural Networks, Natural Language Processing NLP, Optical Character Recognition OCR etc. It works on performing tasks like analysing medical reports, facilitate diagnostic processes and provide actionable insights for healthcare professionals. The step by step working process of the model is described below:

i). **Data Collection and Preprocessing:** The task that is executed first by model is Data collection. The data is collected from various sources like PDF medical reports, CSV reports, input for health chat tool is given in the form of a patient's query. These data are converted into a common format so that it would make it easier for processing the data. Data is extracted from PDF files by utilizing the Optical Character Recognition OCR module, CSV Files are passed in a structured format. Data Cleaning is the process of removing the unwanted data that causes errors in the code, normalize the data by converting the data to a range between 0 and 1. The duplicates present in the data are removed. The missing values are replaced by mean or mode. The steps are performed in order to make the predictions done accurate. Data validation is performed to validate if the data that has been extracted from the reports exist in the expected ranges of the medical parameters. If it does not exist within the range of the data accordingly alerts are given by the model. Medical data is transformed into standardized levels for it to be fed into machine learning models.

ii). **Exploratory Data Analysis:** Exploratory data analysis is used in this tool to understand patterns, spot anomalies present in the data, test hypothesis and check assumptions present in the data. EDA is very useful in the healthcare sector for analysing medical data where understanding trends and patterns in the medical data can lead to better diagnostic accuracy and plan for more effective treatment plans. Time Series analysis is a type of analysis is performed in this tool to monitor the changes of levels of medical parameters over time and its variation is visualized using suitable comparative charts that depict the comparison of the normal parameter levels and actual parameter levels. The correlation analysis is performed to analyse the relationships between various health parameters to derive key insights and predict patient conditions. The outlier detection is performed to identify the extreme or abnormal values of healthcare parameters from medical reports that will have a critical impact in the patient's health based on statistical and machine learning models based analysis.

iii). **Feature Engineering and Model Selection:** To diagnose and predict the patient's health feature engineering plays a crucial role. Here we combine the demographic data, Patient health care reports, medical test results, symptom history and pre-process the data. The Named Entity Recognition feature of the Natural Language processing module is used to differentiate between the Healthcare parameters and Non health related entities that are being extracted from medical reports, test results etc. Auto-ML uses the same machine learning algorithms and same methodology to train and test the models. The difference is that it can optimize the algorithms, automatically test multiple algorithms to get to know the accuracy of predictions made by each model and it chooses the best suited model by choosing the model with highest accuracy score, precision, f1-score, recall score. The Deep learning algorithms like Convolution Neural Networks, VGGNets, ResNet, ImageNet etc are used along with image processing techniques to analyze the scan reports of the patient that are uploaded into the healthcare organ analyzer developed through the user interface. It also utilizes the Auto-ML, NLP tools to coordinate with the analysis where NLP helps the tool in distinguishing between the healthcare parameters and non-healthcare entities, Auto-ML algorithm assists the computer vision algorithm in providing clinical notes like predicting the diseases based on the features extracted by computer vision algorithm, recommendations regarding workouts, diets to be followed, precautions to be taken, etc. The health care tools developed uses Langchain In order to utilize the features of LLMs including GPT-3.5, GPT-4 in the model. It is used in my project to perform clinical notes generation, X Ray report analysis, PDF Analyzer, Healthcare chat tool etc. The Langchain utilizes streamlit, flask for the user interface. The memory module in the LLM integrated by the langchain with the web app makes it remember the context of the interactions with users. Both short term memory and long term memory can be added depending on the use case.

iv). **Model Training and Evaluation:** The models are trained using various machine learning and deep learning algorithms so that it will be exposed to a large number of scenarios, patterns etc. The model is trained based on supervised learning algorithm with the X Ray images and patients diagnostic reports. The Deep learning model CNN's are trained for image analysis based on different types of diagnostic images to classify images. After training the models are evaluated based on various evaluation techniques like accuracy score, f1 score, precision etc. In order to improve the model's performance to the maximum level we perform hyperparameter tuning and model optimization, Now with the usage of the Auto-ML this process is automated.

v). **Interactive Dashboard and Data Visualization:** In this model we use Plotly for visualization of healthcare parameters using various attractive graphs and charts. Exploratory data analysis is performed by using time series analysis when we need to analyze the level of various healthcare parameters over a period of time (say 6 months,1 year, 2 years, 5 years). We perform correlation analysis to find the relationship between various health parameters and we can also find out the patterns, trends, and detect outliers in various health parameters of a patient. The dashboard is developed with interactive colour coded widgets. There are four colours in the dashboard Red, Green, Yellow and Amber. Red is indicating that the parameters is extremely high and the patient is in high risk zone, Amber means the level is moderately high and moderate

risk is there for the patient, Yellow means the level is slightly high and patient is in low risk category and Green means it is in the correct level and there is no risk.

vi). **X-Ray Image Analysis and Prediction:** Image Preprocessing is the process of normalizing the image, converting the image into binary form through binarization i.e. image is converted into 0's and 1's where 0 means dark and 1 means bright. In image preprocessing we can also resize the image, we can change the levels of contrast in the image, brightness of the image also can be changed. Convolution Neural Networks are used to extract features like shape of the organ, size of the organ, impact region of the organ, boundaries of the organ etc by analyzing the medical images. In our model we can analyze the impact region of knees, heart, lungs, ankles. Based on the analysis of the tool the model predicts the impact regions in the organ. The prediction made by diagnosis tool is then sent as a report to the healthcare professional for further diagnosis. This will help to reduce the time required for diagnosis.

vii). **Patient Query Bot and Natural Language Processing:** The Natural Language Processing will first identify the user's intent from the text input and mapping relevant responses a per the user's query. The bot generates the most suitable response for the user's queries by utilizing the LLM's and GPT's capabilities. The user asks some questions related to the disease, medication, duration of the treatment etc. The query will be given to the Question answering tool in the form of prompts and the prompts are converted into numerical format then into vectors called embeddings. These embeddings are divided into chunks (for example : f there are more number of records say 1000 records it will take a lot of time to process so we convert them into batches of 50 or 100 so the processing will be quicker and much accurate). The llms receives these embeddings as inputs and then it researches on the most suitable answer to the query asked by the patient and frames its reply.

viii). **Clinical Decision Support:** The users are given recommendations based on the various KPIs or health parameter's levels in the form of diet, workout, precautions, medications recommendations etc. There is an embedded alert system when the critical parameters in high risk patients show a sudden rise or fall like BP, Diabetes, Pulse etc then the doctors are alerted regarding it through the dashboard app notification.

ix). **Deployment and Monitoring:** The web application developed is deployed with Streamlit interface as the frontend user interface layer and API's are kept at the Backend of the application. Regularly the app is monitored for bugs and issues and are updated regularly.

A) System Architecture

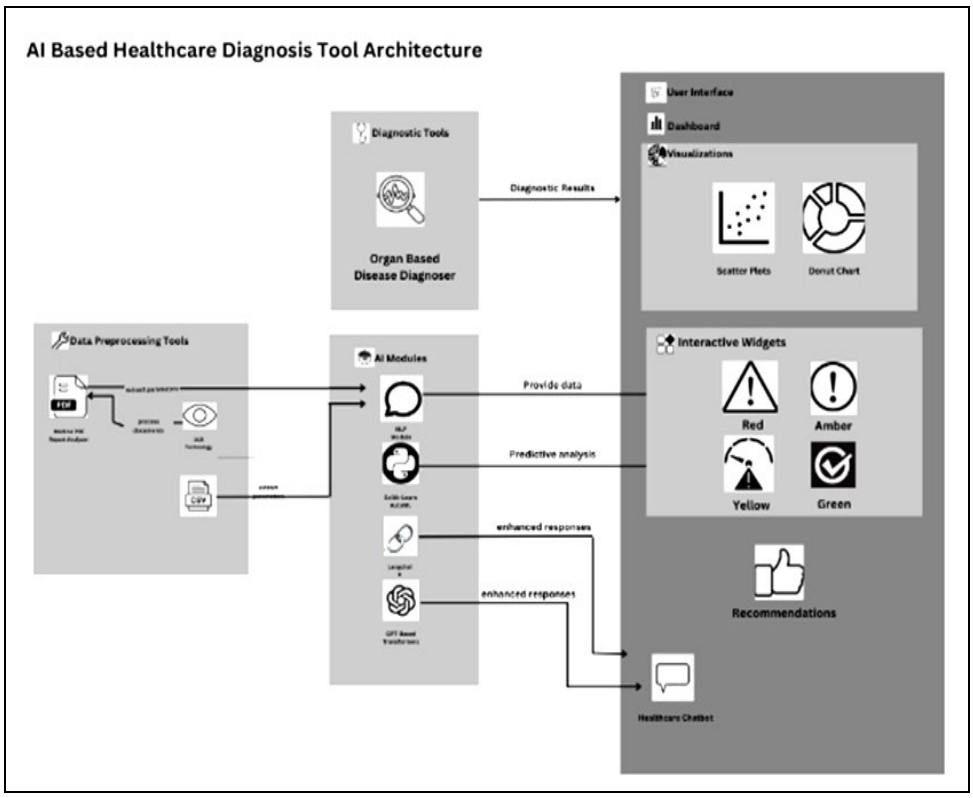

Fig 1: System Architecture Diagram

The Architecture diagram depicted above tell us about the modules which are embedded with the AI-Powered Healthcare Diagnosis tool. The data preprocessing is the first step which is performed by the tool with the Data preprocessing tools like Medical PDF analyser which extracts the content from the PDF by using the OCR Technology. The csv file analyser tools will extract the data from the file by using pandas library. The data that are being extracted by the preprocessing tools are sent to the AI modules that are a part of the Diagnosis tools like NLP, Scikit-learn, Langchain, GPT models which need to perform predictive analysis, provide responses to patients queries that are received as prompts, LLM and langchain enhance the responses quality by using the support of Auto-ML, Transformers and Open-AI. The

dashboard is developed to display the healthcare parameters values depending upon the risk in colour coded widgets where red is for high risk, amber is for moderate risk, yellow is for low risk and green is for no risk. The dashboard also visualizes these parameters with a variety of charts and graphs. Based on the parameter's levels recommendations in the form of what diets to be consumed, workouts can be done, precautions that can be taken, medications can be taken are provided. The organ based analyser system will take the scan report of the patient as input and will predict the impact region and provide some recommendations to the patient regarding diets, precautions etc.

B) Workflow

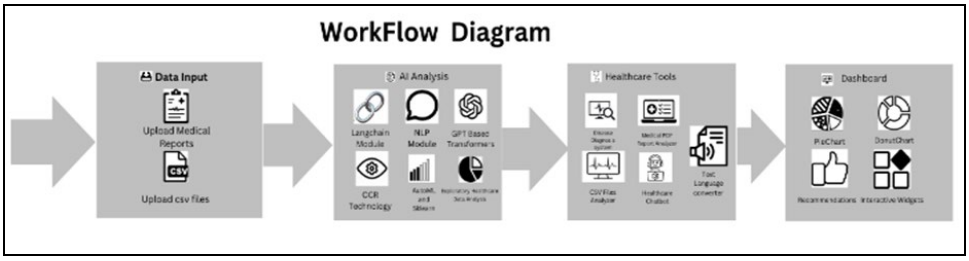

Fig 2: Workflow Diagram

The workflow diagram gives us a clear picture about the basic order in which the functionalities are performed by the AI Based diagnosis tool. Firstly the model will take the input in the form of PDF Files, CSV Files or Image files (for organ analyser). Then the AI Tools Analyses the input by using OCR to extract the content from the pdf files, pandas, numpy and machine learning algorithms to extract content from the csv files. The healthcare tools like PDF Analyzer Healthcare chatbot, CSV Files analyzer etc process the input received by converting the input into series of vectors which are called as embeddings. These vectors are then divided into batches of small numbers as large number of input/vectors in one go will take a lot of time for processing. Then the tool produces a response to the user's query by combining the abilities of Langchain, Auto-ML and LLMs. If the accuracy is not good then it is fine-tuned and algorithms are optimized automatically and processed again. The end point is the visualizations created with attractive colour coded widgets, graphs and charts.

4. Results and Discussions

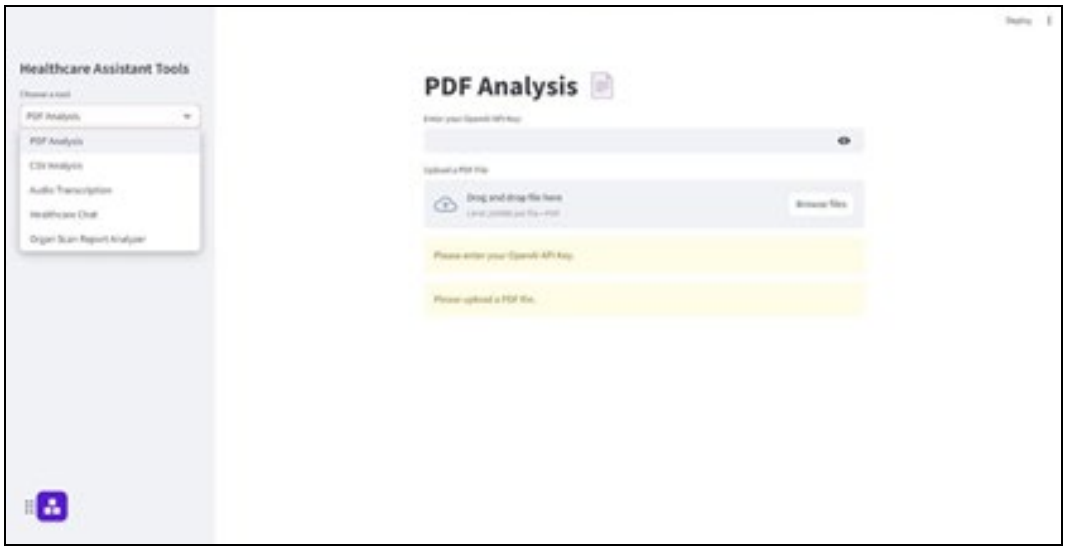

Fig 3: PDF Report Analyzer.

The PDF Analysis tool in Figure 3 have successfully extracted the content from the medical reports and have the ability to reason for queries asked based on the documents.

Fig 4: CSV Analysis Tool

The CSV Analysis tool in Figure 4 have successfully extracted the content from the medical reports in csv format have the ability to reason for queries asked based on the documents.

Fig 5: Audio Transcription Tool

The Audio transcription tool in Figure 5 is converting the doctor's voice that is recorded into text and transcribe it in any other language (For example if doctor speaks in English it can be transcribed in Telugu or Hindi).

Fig 6: Organ Based Impact Analyser Tool.

The organ analyser tool in Figure 6 is analysing the image (scan image) of the patient and predicts the impact regions and provides recommendations for precautions, diets, workout that can be done etc.

Fig 7: Medical Healthcare Chat Application

The healthcare chat app in Figure 7 provides answers to the queries of the patient with a good accuracy level. The Audio transcription tool is used by the doctors to record the audio and convert it into electronic health records based on what is dictated by the doctors in the file.

5. Conclusions

The AI Powered healthcare tool developed by integrating NLP, Auto-ML, Computer Vision leverages the improved level of diagnosis and treatment level for patient. This interactive dashboard developed empowers visualization, colour coded widgets, Real time insights, Alert system and Personalized Recommendations provided by the dashboard. The query processing and question answering mechanism enables improved decision making and patient engagement. These tool also reduces the time and stress levels of the doctors in the hospitals when there is a large number of patients. This tool developed will help in transforming the healthcare sector into a sector with better efficiency and outcomes.

References

1. Crisan D & Fiore M. Enhancing Dashboards with AutoML Capabilities to Improve Result Accuracy. *Journal of Data Visualization and Analytics*. 2021; 8(2):91-108.
2. Donohoe P & Costello G. Data Visualization Literacy and Confidence in Reporting Dashboards. *Journal of Visual Analytics*. 2020; 14(1):47-61.
3. Echeverra L. A Data-Based Approach to Visual Narrativity: Summarizing Data with Critical Insights. *Journal of Visual Communication*. 2018; 9(1):73-88.
4. Elsayed M & Zulkernine M. Security Monitoring as a Service (SMaaS) in Cloud Analytic Applications. *Journal of Cloud Computing*. 2018; 7(2):111-128.
5. Garouni H. An AutoML Model for Processing Big Industrial Data: Auto-Explanation and Analysis. *Industrial Data Science Review*. 2022; 15(2):77-92.
6. Ghavami A. Security Measures in Organizational Dashboards: Protocols, Bug Checking, and Data Encryption. *Journal of Information Security*. 2019; 9(4):300-314.
7. Giovanelli R. Effective Data Preprocessing Techniques in AutoML: Developing Successful Machine Learning Models. *Journal of Machine Learning Research*. 2021; 14(4):233-248.
8. Iyer S. DataScope: A Visual Analytics Dashboard for Large and Multidimensional Datasets. *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics*. 2017; 22(1):234-246.
9. Jacobs R & Rudils T. Security in Data Visualization: Detecting and Reporting Threats during Data Analysis. *IEEE Transactions on Information Forensics and Security*. 2014; 10(2):335-347.
10. Joshi A & Mahavale S. Data Storytelling with Tools like Tableau and Plotly: Guiding Users to Critical Insights. *Journal of Data Narratives*. 2022; 6(3):122-136.
11. Kamarker S. Challenges and Advantages of AutoML Tools in Data Science. *Journal of Data Engineering*. 2021; 10(3):88-102.
12. Khalajzadeh M. Exploratory Data Analysis: Providing Insights in Data-Driven Engineering and Software Development. *Journal of Data Science and Visualization*. 2019; 7(2):105-118.
13. Mahmood T & Aflal M. Embedding Big Data Analytics for Cybersecurity: Enhancing Security in the Digital World. *Journal of Big Data*. 2013; 3(4):205-220.
14. Manatova L. Development of a Vulnerability Management Dashboard for Data Analysis Security. *International Journal of Cyber Security and Digital Forensics*. 2022; 11(3):155-169.
15. Mattmuler J. AutoML for Non-Programmers: Automating Model Selection and Improving Tool Accessibility. *Journal of Machine Learning Automation*. 2023; 12(1):45-59.
16. Nieto F. Data Storytelling Techniques: Visualization, Minimal Text, and Architecture Diagrams. *Journal of Digital Storytelling*. 2022; 13(2):147-163.
17. Obseracher M. Practical Insights for Big Data Analytics through Data Storytelling Techniques. *Journal of Applied Data Science*. 2023; 10(1):202-219.
18. Wu X. DynDash: A Multi-Coordinated Dashboard for Data Visualization. In *ACM Conference on Human-Computer Interaction*. 2017; 18(3):199-208.
19. Chen J, Patel V & Leung W. AI-Driven Diagnostic Models for Organ-Specific Diseases Using Transformers. *Journal of Medical Imaging and Diagnostics*. 2023; 12(3):145-158.
20. Wang L & Kumar A. Exploratory Data Analysis in Healthcare: Real-Time Visualization Techniques. *Healthcare Informatics Research*. 2023; 29(2):89-102.