JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

VISUALIZATION AND PREDICTION OF REAL-TIME SENSOR DATA

Pradnya Talekar Dept. of Electronic and Telecommunication
SIES Graduate School ofTechnology.
Nerul, Navi Mumbai, India

Shubhangi Kharche Dept. of Electronic and Telecommunication SIES Graduate School ofTechnology.
Nerul, Navi Mumbai, India.

Pranjal Thakur Dept. of Electronic and Telecommunication
SIES Graduate School ofTechnology.
Nerul, Navi Mumbai, India

Jinendr Poojary
Dept. of Electronic and TelecommunicationSIES Graduate School of Technology.
Nerul, Navi Mumbai, India.

Rithik Patil
Dept. of Electronic and TelecommunicationSIES Graduate School of Technology.
Nerul, Navi Mumbai, India.

Abstract**—** Nowadays, sensors play an important role in our day-to-day lives. But analyzing the data has always been a vital and precarious part. The work on ‘Data analysis of sensors’ revolves around the concept of analyzing and studying the data obtained from the sensors. The sensors are continuously picking up new data from their surroundings and this data needs to be organized. During this work, real time sensor data was provided by the industry and two goals were achieved. The first one being analyzing the performance of a sensor in a particular period. Each of the cloud devices and the sensors that were used in them were categorized using Power BI software. The second goal was to predict the future problems. Two prediction models were used in the work, namely; Random Forest and Linear Regression model, which fit the data appropriately. Random Forest model gave a mean square error of 4.91% while Linear Regression model gave a mean square error of 3.84%. This analysis was made available to the supervising team by integrating it with our designed website. The analysis available on the website can be accessed by anyone with a valid username and a password and also can get alerts in case something goes wrong. Through this work, we carefully examined and visualized the data procured by the sensor and predicted the future problems. As we used AI/ML models and a visualization software, human error was eliminated while monitoring the sensors. More importantly, it automates typical, boring, and routine jobs that were once completed by humans, which boosts productivity even more, and enhances real-time management of essential machine failure. The two models used for prediction improved the model's effectiveness and produced a mean square error of 0.9%.

I. Introduction

Sensor data analysis involves the collection and analysis of data produced by various sensors to obtain relevant information and insights. Sensors are electronic devices used to measure physical or environmental variables such as temperature, pressure, humidity, light, sound and motion. Data generated by sensors is often collected and stored in databases and then analyzed using various statistical and machine learning techniques to identify patterns, trends, anomalies and correlations. This analysis can help identify relationships between different variables and provide insight into the performance and behavior of the monitored system.

Analysis of sensor data can help identify potential problems before they become critical and optimize maintenance schedules to reduce downtime and costs. In general, sensor data analysis is a critical part of many applications, from industrial process control to environmental monitoring to healthcare. It allows us to gain a deeper understanding of the world around us and make informed decisions based on these insights.

In this work we achieved the following objectives. The first objective was to analyze the data set given based on different parameters like location of sensors, weather, temperature etc. and eliminate the faulty data. The second objective was to identify patterns and trends in the data is one of the primary goals of sensor data analysis. The third objective was to visualize data, the process involves displaying it in a graphical or pictorial style to aid users in understanding its patterns and trends. The fourth objective was to predict the future data, prediction algorithms are applied to sensor data analysis to identify anomalies, forecast future trends, and enhance system performance. And the fifth objective was to develop predictive models, these algorithms can be trained on previous sensor data.

We achieved the objectives in parts. The analysis of data along with eliminating the faulty data was done in Microsoft Excel. Whereas the visualization of the cleaned data was done using Power BI. And for the prediction part we used the AI/ML models which were Random

Forest model and Linear Regression model. We integrated all of this in a website for better access to the supervision team.

The paper is structured to cover Literature Survey in section 2, methodology in section 3, implementation and final results in section 4 and conclusion in section 5, followed by the references in section 6

II. Literature Survey

In the Industry based smart manufacturing environment, machine learning techniques are deployed to identify patterns in live data by creating models using historical data [1]. These models will then predict previously undetectable incidents. The paper [1] initially performs a descriptive statistics and visualization, subsequently issues like classification of data with imbalanced class distribution are addressed. Then several binary classification-based machine learning models are built and trained for predicting production line disruptions, although only logistic regression and artificial neural networks are discussed in detail. Finally, it evaluates the effectiveness of the machine learning models as well as the overall utilization of the manufacturing operation in terms of availability, performance and quality.

Another study [2] proposes a real-time monitoring system that makes use of big data processing, IoT-based sensors, and a hybrid prediction model . First, an IoT-based sensor that gathers data from an accelerometer, gyroscope, humidity sensor, and temperature was created. Real-time, massive, and unstructured kinds are the features of IoT-generated sensor data from the industrial process. Second, to eliminate outlier sensor data and offer fault detection throughout the manufacturing process for the proposed hybrid prediction model, Random Forest classification and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)-based outlier detection, respectively, were employed. The findings demonstrated that the suggested big data processing system and IoT-based sensors are effective enough to monitor the manufacturing process. The suggested system is anticipated to support management by enhancing decision-making and assisting in the reduction of unforeseen losses brought on by manufacturing process flaws.

The paper [3] focuses on sensor data analysis along with anomaly detection specific to the process sector because the placement and nature of the data generated from these sensors follows a specific pattern during process flow. This data is more structured than other types of big data, so this paper presents a generic framework with an ensemble of methods such as probability and statistics, Neural Networks and Clustering. But unseen data is wrongly predictable by Neural nets so clustering is used as an Unsupervised learning model and eventually improves prediction and anomaly detection accuracy of equipment as well as process flows. No single framework is available to fully analyze sensor data stream related to independent, correlation based, group wise with respect to process flow segmentation and process and sub process hierarchy analysis.

A model is proposed with the help of machine learning [4] that will be used in multiple chambers fridges to keep indicating the time remaining for the inner temperature to go beyond the allowed range, and if the time is short, the system will propose to the pharmacist not to open that particular room and proposes a room that has enough time slots (time to reach the upper limit temperature). By using training data obtained from a thermoelectric cooler-based fridge, we constructed a multiple linear regression model that can predict the required time for a given room to reach the cut-off temperature in case that room is opened. The built model was evaluated using the coefficient of determination R2 and is found to be 77%, and then it can be used to develop a multiple room smart fridge for efficiently storing highly sensitive medical products.

The paper [5] covers the source of data acquisition, highlights the technical features of the data source, and explains the data collection procedures and requirements using a big data analysis methodology. Establish a data analysis system, define the aim of big data analysis, and explain the data analysis procedure. Data collection, storage, analysis, and mining are the four processes in big data analysis. The two components of data acquisition are acquisition and preprocessing, which is just narrow data acquisition. The Internet of Things (IoT) data analysis process involves planning and systematically gathering data, processing data, and conducting data analysis to produce information. The algorithm's support and execution processes are composed of the analysis process. To increase the efficiency and logic of data processing, data analysis should be correctly used throughout the whole data processing cycle, from data gathering in the field perception layer through data transmission in the communication layer to data processing in the application layer.

In recent years, there has been a lot of interest in widely disseminating data analysis knowledge online rather than keeping data and analysis skills in-house. Among stakeholders, such as data owners, users, and analysts, a data exchange platform is a specific kind of digital platform that facilitates data exchange. The data suppliers, however, separately obtain and retain the datasets used on such platforms for their own purposes.

There is currently little information available to examine the systematic organization and combination of these datasets because they are not founded on the principles of coordination and combination. A study [6] concentrates on metadata, data summaries, and analyzing data similarity on a platform for data exchange using natural language processing. We utilize the metadata from the data exchange platform Kaggle in our experiments. Our method translates data descriptions to vectors and uses the vectorized methods word2vec and BERT to compare how similar the data is. Then, by calculating the cosine similarity between each vector, our method determines the distances between each vector. Based on the findings of our experiments, Kaggle shares the same characteristics as other data exchange platforms. Additionally, the outcomes demonstrated the viability of the strategy for extracting similar data pairs based on natural language processing.

The era of big data has arrived as a result of the ongoing advancement of contemporary Internet and e-commerce technologies, which has caused network data to rise geometrically [7]. Many industry resources are contained in the network's huge data, so how to correctly mine that information has become a new problem for businesses. The traditional Apriori mining algorithm and the most recent Hadoop technology are used in this research to build a new data mining and analysis technology. The two aspects below are the key areas of attention for this technology. The first step is the upgrade of the conventional method, which mostly uses Hadoop technology; the second step is the processing of mining data in parallel for analysis. The new large data mining analysis technology, which offers some references for the existing logical data mining and network data applications, is made up of the two main components.

Pervasive sensing is one of the most prominent technologies being adapted by the current process industry. Every process industry is highly equipped with wireless sensors for process monitoring in which location, human intervention is to be limited. Thus, a major challenge with these numerous sensors is to store and analyze large volumes of sensor data stream. The paper [8] focuses on sensor data analysis along with anomaly detection specific to the process sector because the placement and nature of the data generated from these sensors follows a specific pattern during process flow. This data is more structured than other types of big data, in which data is more unstructured. No assurance that any single algorithm can produce optimized results. So this paper presents a generic framework with an ensemble of methods such as probability and statistics, Neural Networks and Clustering. Here Neural Net is a supervised learning model to predict new data based on trained data. But unseen data is wrongly predictable by Neural nets. For that reason clustering is used as an Unsupervised learning model to efficiently handle concept drifts in sensor data streams. These solutions are implemented to various data scenarios with practical means to improve prediction and anomaly detection accuracy of equipment as well as process flows. To the best of our knowledge no single framework is available to fully analyze sensor data stream related to independent, correlation based, group wise with respect to process flow segmentation and process and sub process hierarchy analysis.

As the big data Era, the amount of data in education shows explosive growth. Big data analysis system architecture intends to provide a reference for large educational data analysis. Through literature analysis and network investigation, the paper [9] reviewed the existing mature system of large educational data analysis systems. From two aspects of generality and difference to compare, this paper proposed a research framework from several aspects, including analysis of the common thinking, the generality of open source thought, analysis of the common areas, technology framework, the core technology and special functions. This paper compares three mainstream data analysis systems, summarizes the enlightenment for the analysis of large data, puts forward a kind of intelligent analysis system of education big data, and sums up the development prospect of big data analysis systems.

The project seeks to provide a data gathering methodology and data mining algorithms for an information system for monitoring the health of pupils. Using artificial intelligence techniques, a number of issues relating to analysis, pattern search, result visualization, and a mix of medical, social, and academic data are resolved within the context of this scientific challenge. Particular emphasis will be placed on Kazakh educational institutions in this study. A variety of internal and external stakeholders can use an information system with visualization of the outcomes of data gathering and analysis on health-related indicators as a regularly updated information resource for the development of various social or medical assistance programs. Based on specifically gathered information from medical specialist consultations, the paper [10] examined and evaluated machine learning methods. Additionally, a statistical analysis was done.

In the work presented in this paper, we analyzed and visualized the data using Power BI software. Additionally we used Random Forest and Linear Regression models for prediction of future problems. The analysis was made available to the supervising team by integrating it with our designed website.

III. METHODOLOGY

The data for the proposed work was collected in real time from the sensors deployed at various locations in India. Permission was obtained from Industry experts to work on that data. We collected, cleaned and tried to visualize the data, and find patterns from it. After that an algorithm was applied to the training data set to train the model. And then the testing data set was applied to the model. A block diagram is seen in Fig. 1 which represents how the analysis of data was done. Once the model was ready, it helped us to observe the performance of a sensor in a particular period of time and predict the future problems.

The first step in achieving the objective was to clean the data by removing redundant factors. As we were dealing with a lot of sensors, we had to differentiate the data for each sensor separately and compile it together to avoid the confusion. We did this using Microsoft Excel.

The second step was to visualize the data which refers to the process of presenting data in a graphical or visual format. To make complex data sets easier for consumers to grasp and interpret, interactive visualizations are made in the form of charts, graphs, maps, and other visuals. The main objective of data visualization is to efficiently and effectively convey information so that users can obtain new perspectives, spot trends, and come to wise judgments. Data is made more comprehensible and accessible by being presented visually, especially for non-technical or novice users. A variety of choices for creating visual representations of data are offered by data visualization tools and software, including Tableau, Power BI, Python libraries like Matplotlib and Seaborn, or JavaScript libraries like D3.js. Depending on the type of data, the goal of the study, and the intended audience, these tools provide flexibility in selecting the best visualization strategies.

Visualization aspect was important because it helps study the pattern that the data is following. We used Microsoft Power BI software, a tool for business intelligence and data visualization where each cloud device and the sensors used in it are classified. The sensor data for each minute has been visualized

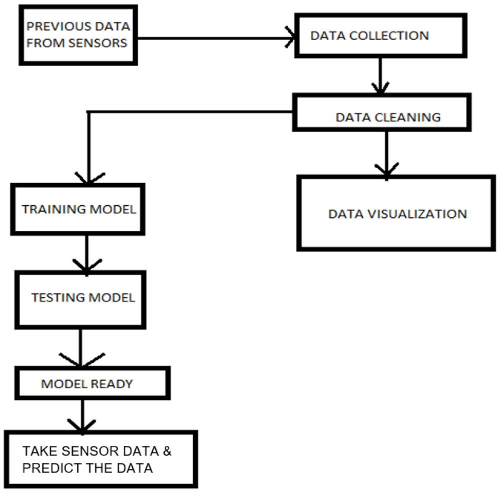

Fig 1 : Block Diagram data visualization and prediction

.

Following data cleansing and visualization, we moved on to the prediction phase.We used AI/ML models for the prediction process. Some of the models that we used were the autoregressive model, ARIMA (Autoregressive integrated moving average) model which is a well-known and effective time series forecasting model that may be used to estimate future values using the dependencies and patterns in historical data and SARIMA (Seasonal ARIMA) model which is an addition to the ARIMA model that takes seasonality into account when predicting time series. SARIMA is intended to identify and model patterns that recur in the data on a daily, weekly, or yearly basis, such as seasonality.. But none of these models gave satisfactory results.

Other models that we used and produced good outcomes were Random Forest model and Linear Regression model which fit the data appropriately. In both classification and regression tasks, the well-liked machine learning method Random Forest is employed. It used ensemble learning to blend different decision trees to predict the future. With each tree trained on a different random subset of the data and features, the algorithm created a forest of decision trees. Although Random Forest is a strong and adaptable algorithm, it might not always be the ideal option for every problem. Before choosing the best model for your prediction task, we decided to compare various methods and take into account the unique properties of your dataset. A useful baseline of performance, interpretability, and insights were supplied by a linear regression model for more sophisticated models. However, it's crucial to remember that linear regression makes certain assumptions about the data, such as their linearity, and that its success depends on the specifics of the problem at hand. These two models obtained the mean square error of 0.9 and increased the efficiency of the model.

IV. RESULT & DISCUSSION

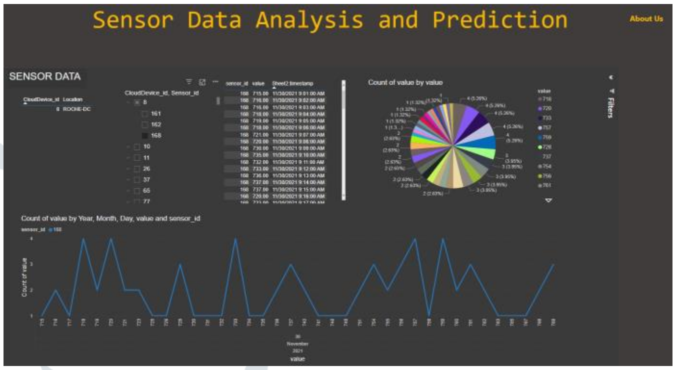
Sensor Data Analysis and Prediction
About Us
SENSOR DATA
CloudDevice_id, Sensor_id
Count of value by value
Count of value by Year, Month, Day, value and sensor_id
Filters
value

Fig 2 : Visualization of sensor id 168

In Fig. 2, visualization of data is shown. We observe the cloud device ids and the sensor ids within it. On top right of the diagram, a pie chart is shown where how many times the count value is repeated is displayed. The same data is shown in the form of a line graph for sensor id 168. In the graph, we have represented day, month, year and value on x-axis. While on the y-axis, count of value is represented.

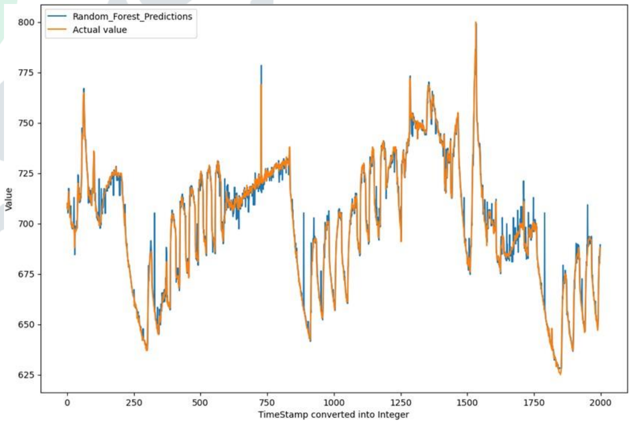

Fig 3 : Prediction using Random Forest model

The graph in Fig. 3 represents Random Forest and shows the predicted values with actual values. The x-axis represents a timestamp which has been converted into an integer while the y-axis represents the actual value of the sensor. From the graph it is understood that predicted values are very close to the actual values. This means the data fits the model accurately and gives proper predictions. Mean Squared Error for Random Forest Model for sensor id 168 is 4.92%.

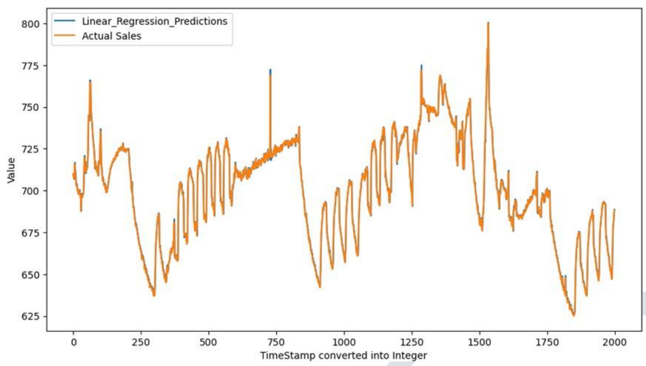

Fig 4 : Prediction using Linear regression model

The graph in Fig. 4 represents the Linear Regression Model, and shows the predicted values with actual values. The x-axis represents a timestamp which has been converted into an integer while the y-axis represents the actual value of the sensor. From the graph it is understood that predicted values are very close to the actual values. This means the data fits the model accurately and gives proper predictions. Mean Squared Error for Linear Regression Model for sensor id 168 is 3.8470615803522374%.

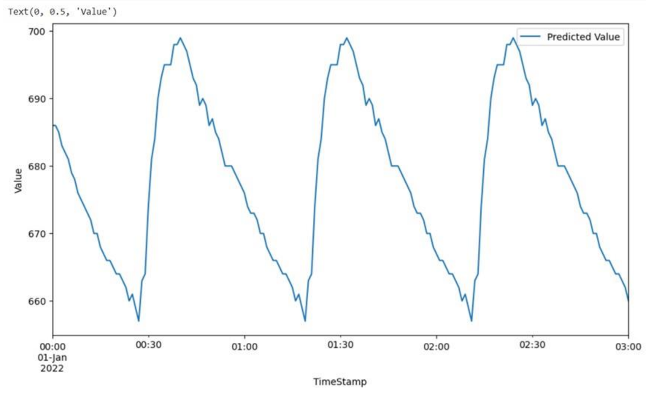
Text(0, 0.5, 'Value')

Fig 5: Future Prediction of values

The above graph in Fig. 5 represents future predictions in the values of sensors. The x-axis represents timestamp while the y-axis represents the value of sensors. It is seen that around 00:30 the value of the sensor is less than 660 and at 1:00 the value of the sensor goes up to 674. A pattern is followed throughout the graph where the values keep increasing till they reach almost 700 and after that the values start decreasing.

V. CONCLUSION

The suggested system focused on data analysis based on many aspects, such as weather, and is predicted to identify incorrect data and further eliminate it. With the use of a clear methodology, future data can be forecasted based on historical data. Since this paper clearly demonstrates a real-time application, data analysis from sensors can be extremely important to corporate operations. The sensors continuously collected fresh data from their surroundings, which was then carefully organized. All objectives which we set to achieve were accomplished. The first one was to examine a sensor's performance throughout a specific time period. Microsoft Power BI software was utilized to classify each cloud device as well as the sensors that were housed inside it.. The following step is to foresee potential issues in the future. Linear Regression and the Random Forest models, which both adequately fit the data, were the prediction models that we utilized. These two models improved the model's effectiveness and produced a mean square error of 0.9%. Sensor analytics can be considered as a future investment because IoT devices will soon play an even larger part in our lives. Additionally, in order to reduce costs and boost revenues, businesses strive to be more efficient and intelligent in their operations. In smart houses, sensors are used to track a variety of factors, including temperature, humidity, and energy usage. Home owners can optimize their energy use, increase home security, and find possible issues like water leaks or mold growth with the help of data analysis from these sensors.In the agricultural sector, sensors are used to track crop development, meteorological conditions, and soil moisture levels. These sensor data processing can assist farmers in waste reduction, crop production prediction, and irrigation optimization. The importance of sensor data analytics to corporate operations is a result of all these reasons, which make it crucial to invest in an analytics platform that can stream and analyze data in real-time from various sensors. Sensing data analysis can and will change the way businesses function by cutting costs, being more efficient, and ultimately boosting profits. Privacy of the data available on the website can be further improved in the future work. Further, instant suggestions can be made if any abrupt changes occur in the system.

VI. REFERENCES

[1] Alexandra Moraru, Marko Pesko, Maria Porcius, Carolina Fortuna, Dunja Madenic, Using Machine Learning on Sensor Data, Journal of Computing and Information Technology - CIT 18, 2010, 4, 341–347 doi:10.2498/cit.1001913

[2] Muhammad Syfraduin, Ganjar Alfian, Norma Latif Fitriyani, Jongta Rhee, Performance Analysis of IoT-Based Sensor, Big Data Processing, and Machine Learning Model for Real-Time Monitoring System in Automotive Manufacturing, 2018 Advances in Sensing, Processing and Transmission for IoT-Oriented Sensors

Networks, Sensors 2018, 18(9), 2946; https://doi.org/10.3390/s18092946

[3] U. Surya Kameswari,Prof. I. Ramesh Babu,Sensor Data Analysis and Anomaly Detection using Predictive Analytics for Process Industries, 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI).

[4] Joseph Habiyaremye, Marco Zennaro, Chomora Mikeka, Emmanuel Masabo,A Data-Driven Predictive Machine Learning Model for Efficiently Storing Temperature-Sensitive Medical Products, Such as Vaccines: Case Study: Pharmacies in Rwanda, 2021 Journal of Healthcare Engineering

[5] H. Li, "Research on Big Data Analysis Data Acquisition and Data Analysis," 2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), Xi'an, China, 2021, pp. 162-165, doi: 10.1109/CAIBDA53561.2021.00041.

[6] H. Sakaji, T. Hayashi, Y. Fukami, T. Shimizu, H. Matsushima and K. Izumi, "Retrieving of Data Similarity using Metadata on a Data Analysis Competition Platform," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 3480-3485, doi: 10.1109/BigData52589.2021.9671414.

[7] Z. Dong, "Research of Big Data Information Mining and Analysis : Technology Based on Hadoop Technology," 2022 International Conference on Big Data, Information and Computer Network (BDICN), Sanya, China, 2022, pp. 173-176, doi: 10.1109/BDICN55575.2022.00041.

[8] U. S. Kameswari and I. R. Babu, "Sensor data analysis and anomaly detection using predictive analytics for process industries," 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI), Kanpur, India, 2015, pp. 1-8, doi: 10.1109/WCI.2015.7495528.

[9] J. Chen et al., "Research on architecture of education big data analysis system," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 2017, pp. 601-605, doi: 10.1109/ICBDA.2017.8078706.

[10] M. Mansurova, M. Zubairova, N. Kadyrbek, G. Tyulepberdinova and T. Sarsembayeva, "Data Analysis for The Student Health Digital Profile," 2021 16th International Conference on Electronics Computer and Computation (ICECCO), Kaskelen, Kazakhstan, 2021, pp. 1-6, doi: 10.1109/ICECCO53203.2021.9663804.