LEVERAGING DATA ANALYTICS AND LSTM MODELS FOR PREDICTIVE MANAGEMENT IN HOSPITALS

Bao-An Nguyen[1*], Minh-Cuong Nguyen[2], Duong Ngoc Van Khanh[3]

Abstract – *Current hospital management struggles to adapt to fluctuating patient volumes, leading to inefficiencies and crowding. Traditional paper or spreadsheet-based methods lack real-time responsiveness. This research proposes a novel approach utilizing data analytics and prediction using long short-term memory models to modernize medical processes and optimize patient experience. The study aims to develop a data-driven system that aggregates and visualizes information on patient numbers, medical visits, and clinic status. A data warehouse information system was built by leveraging large datasets from Tien Giang Provincial General Hospital (Vietnam). Data analytics tools and long short-term memory models were then employed to analyze trends and predict future patient volumes and disease patterns. This system offers several advantages: improved scheduling, regulated patient flow, optimized resource allocation, and a more convenient and efficient medical experience. It not only empowers managers with real-time insights and predictive capabilities but also paves the way for broader applications of data analytics and prediction models in healthcare management.*

Keywords: data analytics in healthcare, data visualization time series prediction, data warehouse, long short-term memory.

I. INTRODUCTION

As the demand for medical services continues to rise, hospitals worldwide face serious overload, impeding healthcare delivery and compromising patient safety [1]. Maintaining efficiency and organization in medical procedures is a significant challenge for busy hospitals that serve thousands of patients daily. This necessitates modern strategies such as extending appointment hours, implementing flexible treatment plans, optimizing human resources, and accurately forecasting patient populations.

However, traditional methods are inadequate to manage this increased demand. Current medical management systems, for instance, often suffer from inefficiencies due to their limited flexibility and data processing delays. The daily influx of patients generates vast amounts of medical data that need processing and storage. Unfortunately, existing systems often rely on fragmented relational databases with numerous separate tables, making it cumbersome and time-consuming to generate statistics and evaluate information. Furthermore, retrieving and exporting data is restricted due to the need for complex data structure understanding and frequent table joins. These factors lead to slow, cumbersome queries, hindering system performance and data accessibility.

Utilizing data analytics to visualize and predict medical data can enhance hospital management and elevate patient experience [2]. This article proposes applying data analytics to visualize and predict medical examination and treatment data. Interactive charts, graphs, and maps are used for data visualization, transforming complex medical information into an intuitive and easily digestible format for informed decision-making. Additionally, a clinical data prediction model leveraging machine learning techniques like long short-term memory (LSTM) neural networks is trained on historical data to forecast future outcomes. This empowers more effective resource management and improves the planning and implementation

[1,3]Tra Vinh University, Vietnam
[2]Master Student, Tra Vinh University, Vietnam
*Corresponding author: annb@tvu.edu.vn
Received date: 28th June 2024; Revised date: 24th July 2024; Accepted date: 22nd August 2024

of medical examinations and treatment protocols.

This project aims to enhance the quality of medical services by providing a comprehensive view of medical data and facilitating informed decision-making. Using data from Tien Giang Provincial General Hospital, data visualization and prediction were employed to reveal insights hidden in management databases. By analyzing this data, the system can predict patient visiting trends, identify overloaded clinics, and provide actionable information to optimize medical resources. The benefits of this research are not limited to effective hospital management but also extend to improving the ability to assess public health, predict and prevent overcrowding, and optimize resources. medical force. In this way, data analytics is more than just a management support tool but also the key to opening up new possibilities in healthcare and hospital management.

II. BACKGROUND AND RELATED WORKS

A. Data analytics in healthcare systems

Data analytics plays a crucial role in understanding public health trends. It empowers us not only to analyze the current situation but also to predict future outbreaks and make informed decisions to optimize care processes [3]. The infrastructure of data analytics is often built on data warehouses using the star/snowflake schema models and the ETL (extract, transform, load) process [4]. Pre-computed measures stored in data warehouses serves as the foundation for various applications such as data analysis, reporting, and even disease prediction. Consequently, data warehouses are not simple storage; they ensure data accuracy and efficiency through proper processing, transforming them into a valuable source of high-quality information.

A key technique within data analytics is Online Analytical Processing (OLAP). OLAP enables multidimensional data analysis, allowing users to drill down into specifics, roll up for broader insights, and perform various 'slice and dice' operations for a comprehensive and flexible view [5]. This online, interactive tool empowers users to perform multidimensional analysis directly on the data warehouse, facilitating quick decision-making based on these insights to enhance the quality of medical services. As a result, OLAP has become a valuable method for data analysis in healthcare systems [6].

By combining data analytics, data warehousing and OLAP, this research delves into the detailed and comprehensive application of these tools in healthcare management (Figure 1). This approach promises to not only predict clinical visit amount but also optimize care processes and support informed decisions, ultimately leading to improved medical service quality.

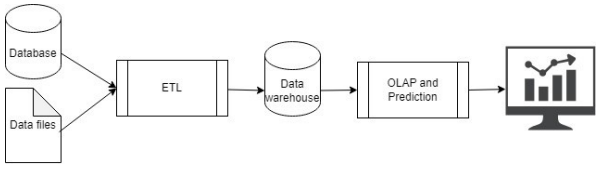

Fig. 1: Data analytics with data warehouse and dashboard

B. LSTM prediction for time series

In healthcare information, seasonal and non-seasonal trends often emerge, revealing health events occurring within the community. Understanding and analyzing these trends provides valuable insights for healthcare professionals in management and care planning. Time series analysis is a powerful tool in medicine, allowing continuous observation over time to analyze medical variables such as patient numbers or exam frequencies, thereby aiding in the prediction and optimization of healthcare resources [7]. One particularly effective approach for time series prediction in medicine is the LSTM model [8].

LSTM networks excel at learning and remembering past information, making them ideal for analyzing medical data with long-term dependencies. This capability stems from their unique internal structure. Within the model, a crucial component is the LSTM layer [9]. Designed to address challenges with long-term information

retention in machine learning, it utilizes three key gates: the Forget gate, Input gate, and Output gate. These gates function as control mechanisms, allowing the LSTM to automatically learn what information is most relevant. The Forget gate decides what past information to discard, the Input gate determines what new information to remember, and the Output gate controls what information is passed on to the next stage of the network. The LSTM process can be broken down into three key steps as depicted in Figure 2.

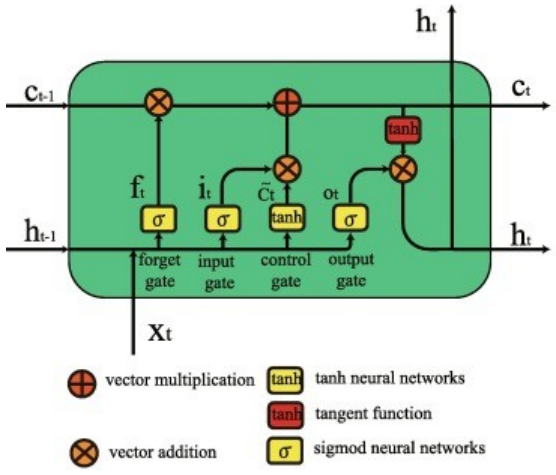

Fig. 2: LSTM network architecture [10]

i) Selective remembering (Forget gate): The network first considers past information from the cell state. A 'Forget gate' analyzes this data, deciding what elements are still relevant and what can be discarded. This prevents the network from becoming overwhelmed with irrelevant historical information.

ii) Learning new information (Input gate): The network then focuses on the current input data. An 'Input gate' determines what new information is crucial and updates the cell state accordingly. This ensures the network incorporates the latest trends into its memory.

iii) Information flow (Output gate): Finally, the network decides what information from the updated cell state is most valuable for future predictions. The 'Output gate' controls the flow of this information to the next stage of the network, effectively determining the short-term memory used for analysis.

The interplay between the cell state (long-term memory) and the hidden state (short-term memory) allows LSTMs to retain crucial information across extended periods. This ability to learn and remember past trends is critical for accurate predictions and analysis in time series environments, making LSTMs a powerful tool for various machine learning applications [11]. Applying LSTM models to medical time series data allows for accurate predictions about future medical situations [11–13]. These forecasts not only provide a deeper understanding of trends but also create diverse and flexible scenarios for future planning.

III. METHODOLOGY

Beside the development of a visualization dashboard of clinical visits data using data warehouse and OLAP, the study also aims to leverage the power of time series analysis and LSTM models to predict number of upcoming patients in every clinical unit. By applying these methods to real-world healthcare data, the study expects to generate detailed and comprehensive insights. These insights could then be used to support decision-making and improve management within the complex landscape of modern healthcare.

A. Data warehouse and visualization

To enable the visualization dashboard, the infrastructure was built for using the following components: i) Data warehouse: Built on Microsoft SQL Server 2017 for data management and storage; ii) Data analysis and visualization services: Microsoft Power BI Desktop is used for research, designing visual structures, and implementing OLAP analysis; iii) Prediction service: LSTM prediction is implemented in Python fetching predicted data to the visualization board.

Data warehouse construction

- Data modeling

The data warehouse employs a star schema model (Figure 3) for optimized query performance, simplified information discovery, and

seamless integration of data from various sources. The model consists of: i) Fact table (Fact Examinations): Stores data on examination counts, results, time, clinic, symptoms, disease types, and disease information; ii) Dimension tables: Provide details on patients, clinics, diseases, and time periods. Foreign keys link these tables, creating a comprehensive data storage system for extensive healthcare analysis and informed decision-making.

- ETL process

Data is extracted from the hospital's existing database, transformed (normalized), and loaded into the data warehouse tables. This involves: i) Creation of a new data system: Based on the designed schema model and linked to the hospital's database; ii) ETL automation: Stored procedures in SQL Server ensure consistent and compatible data extraction and loading; and iii) Job scheduling: SQL Server Agent's Jobs function automates weekly or monthly ETL processes. This ensures efficient and reliable data warehouse deployment and maintenance. The ETL process is designed for continuous data updates, guaranteeing that the data warehouse remains current for healthcare analytics and prediction activities.

Fig. 3: Star schema of the data warehouse

Data visualization with OLAP and PowerBI

Power BI Desktop connects to the data warehouse on the SQL Server to perform OLAP analysis across various dimensions like time, disease type, and patient demographics. OLAP cubes are created to perform calculations and store required values. For example, a cube might calculate the number of examinations by clinic and date (Figure 6). The system offers a user-friendly interface for exploring healthcare data through the following visualizations:

- Informative charts and graphs: Column charts, line charts, and pie charts are utilized to present data in clear and concise formats. These visualizations effectively depict examination volume trends, disease type distribution, and patient examination details (Figure 7).

- Interactive dashboards: Powerful dashboards empower users to filter data by year, month, quarter, and specialty. These selections dynamically update the visualizations in real-time, enabling multidimensional analysis. Users can compare data across various factors simultaneously, gaining a more comprehensive understanding of healthcare trends.

To gain insights into healthcare trends, this study analyzed data from the dimensional and fact tables, including:

- Number of patients by examination stage
- Top five clinics with the highest and lowest clinical visit amounts
- Disease and disease categories
- Number of patients by demographics and time period
- Gender distribution in examinations (including top diseases and clinics)
- Identifying common symptoms and diagnoses
- Visualizing examination volume data across various timeframes
- Creating data visualization pages for disease types, groups, and patient information.

By combining Power BI and OLAP techniques, this system creates visual models that empower medical professionals and administrators to gain deeper insights into medical examination and

treatment data. This ultimately supports strategic decision-making, enhances healthcare service quality, and enables a flexible response to public health challenges.

B. Building an LSTM prediction model for clinical visit amount

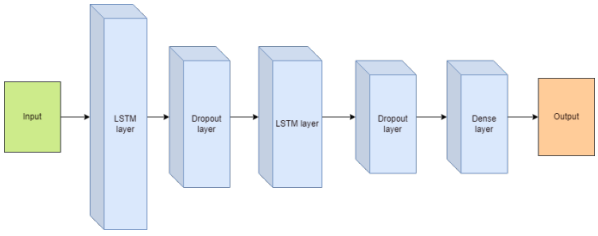

Fig. 4: Architecture of the LSTM model used in this study

IV. RESEARCH RESULTS

The system utilizes a computer with a Core i5 processor and 16GB RAM to build the system. PowerBI enables the dashboard with visual charts and functional charts to support the management and analysis of clinical visit data, based on the data connection to the MS SQL Server data warehouse. Data cleaning, transformation, and conversion to date time format for time-based analysis were performed prediction. Python toolkits for data wrangling and machine learning, Pandas, Tensorflow, and Keras, were used to implement the data processing and prediction tasks.

A. Data acquisition

Historical data on patient visits (1,866,757 visits from 2017 to 2022) from Tien Giang Provincial General Hospital was collected. The data warehouse integrates management data, including:

- Patient Records (384,006): name (họ tên), year of birth (năm sinh), gender (giới tính), address (địa chỉ).

- Medical examinations (2,998,442): dates (ngày), types (loại), symptoms (triệu chứng), clinics (phòng khám), diagnoses (chẩn đoán), and treatment direction (hướng điều trị) (covering 6 years).

- ICD Codes (24,652): ID, names (tên bệnh), and disease groups (nhóm bệnh) (aligned with regulations from Ministry of Health of Vietnam).

- Time data: Allows multidimensional views of disease development (daily (ngày), monthly (tháng), quarterly (quý), yearly (năm)).

- Clinic information (79 clinics): names (tên phòng khám), types (loại), and departments (khoa). To ensure patient confidentiality, the data underwent de-identification before extraction. Time series for daily, monthly, or weekly examination trends were created. Data was grouped by clinic and year-month to calculate examination counts (as depicted in Table 1). Data sequences suitable for the LSTM model were prepared. The data was divided into training (80%, 2017–2021) and testing sets (20%, 2021–2022) for model training and evaluation.

Table 1: Training data samples

ID	ICD_id	MonthYear	visit_count
4080	112162	2022-06	48
4081	112162	2022-07	41
4082	112162	2022-08	123
4083	112162	2022-09	57
4084	112162	2022-10	2

B. Model training

The model was implemented on the same computational hardware as mentioned. Python 3.8.0 with libraries like Matplotlib, Pandas, TensorFlow were used for building the model.

To optimize the LSTM model architecture for clinical visit amount prediction, a series of experiments was by varying the number of LSTM layers and the number of units within each layer. As shown in Table 2, both parameters significantly impacted model performance.

Interestingly, the research observed a counter-intuitive trend: increasing the number of layers led to a rise in error. Therefore, a 2-layer LSTM architecture achieved the best prediction accuracy. Similarly, the number of units per layer

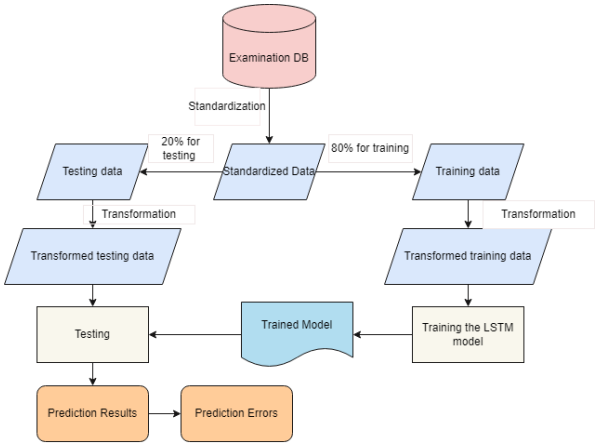

Fig. 5: Workflow of training and testing the LSTM model

Table 2: Error rates with varying LSTM layers

Number of LSTM layers	MAE	RMSE
2	**126.2**	**187.45**
3	128.99	188.87
4	130.86	190.68
5	131.89	192.17
6	130.32	189.26

played a crucial role. The optimal value of 1024 units per layer through experimentation was identified (refer to Table 3 for details).

Table 3: Error rates with varying units in each layer (2 LSTM layers)

Number of units in 1st LSTM layer	Number of units in 2nd LSTM layer	MAE	RMSE
2048	1024	126.21	189.73
1024	**1024**	**124.43**	**187.39**
1024	512	125.94	189.7
512	512	134.97	193.59
512	256	127.18	192.29
256	256	126.67	186.82
256	128	134.39	193.68
128	128	130.05	191.05
128	64	132.11	193.37
64	64	129.84	190.16

Finally, the optimal LSTM configuration consisted of two layers, each containing 1024 units (Figure 5). This architecture resulted in errors of 126 medical examinations per clinic per month and 124 clinic visits per month, respectively (refer to specific prediction scenarios). This optimized model demonstrates promising accuracy and effectiveness, providing a valuable foundation for analyzing and forecasting clinical visit amounts in this study. The selected model was trained using the Adam optimizer and the mean absolute error (MAE) loss function to minimize the difference between predicted and actual values. The training ran for 100 epochs with a batch size of 16. These settings provided a good balance between training efficiency and model performance (Table 4).

Table 4: Evaluation of training parameters: epochs & batch size

Epochs	Batch size	MAE	RMSE
50	128	132.73	194.76
100	128	131.01	192.165
150	128	129.39	191.75
200	128	136.11	195.06
300	128	127.48	188.79

C. Model evaluation

The trained model was put to work generating predictions for unseen data. A new column was added to the data table to store these predicted values. To assess the model's accuracy, the test set was used to calculate root mean squared error (RMSE) and MAE. These metrics measure the difference between predicted and actual values. Lower values indicate a closer match between predictions and reality. Finally, both the predicted values and the actual data were saved to a CSV (Comma-Separated Values) file for further analysis and visualization in Power BI.

The model achieved impressive results in predicting daily clinic volume for 25 clinics over a 2-year period. One-day predictions had a low error of only eight visits per clinic per day (MAE = 7.95, and RMSE = 12.24). For the predicting problem of monthly patient numbers, the model achieved high prediction accuracy for monthly

patient numbers, with low MAE and RMSE values of 24.9 and 28.2 (visits per clinic per month). This indicates a close match between predicted and actual data, making the results suitable for visualization on the dashboard.

D. Data visualization

Aggregated data from the data warehouse, along with predictive data generated by LSTM models, can be effectively visualized through dashboards to offer valuable insights for decision-making support. This subsection presents various methods of data visualization using charts. It is important to note that all text in the dashboard is presented in Vietnamese as per user requirements. For translations of the Vietnamese terms, please refer to figure notes.

Visualization of clinical data

This function visualizes data on patient visits over days, weeks, or months, using column or line graphs. This provides a clear view of patient volume trends and fluctuations over time (Figure 6). Users can easily perceive daily, weekly, or monthly changes, aiding in smart decision-making and effective prediction. This is particularly valuable for resource planning and meeting clinical needs, making data visualization an essential tool for resource management and improving medical clinic performance.

Visualization of patient data

This function visualizes patient information, including age, gender, and detailed medical data, using pie charts, bar charts, or other visual representations. These tools enable users to easily understand the characteristics and distribution of patients within the hospital management system. Data visualization provides an overview of key factors such as age and gender distribution, enhancing the understanding of patient populations (Figure 7). This insight supports decision-making in health policy and service optimization, playing a significant role in managing information and improving the health system's response to patient diversity.

Visualization of disease volume by category

This function visualizes the number of diseases by type or group, using tools such as word

Fig. 6: Visualization of clinical visit amount by clinic and time

Note: Lượng Bệnh Đến Khám Theo Thời Gian = Number of Patients Visiting for Examination Over Time; 5 Phòng khám Nhiều/Ít Lượt Khám Nhất = Top 5 Clinics with the Most/Least Visits; Tỉ lệ đối tượng khám bệnh = Proportion of Patient Groups; Số Bệnh/Phòng khám = Total Patients/Clinics; Tổng Lượt Khám Bệnh = Total Number of Visits; Lượt bệnh khám theo từng tháng = Number of patients examined per month; Lượt Khám Theo Chuyên Khoa = Number of Examinations by Specialty.

Fig. 7: Visualization of patient amount by year of birth and gender

Note: Lượng Bệnh Theo Người Bệnh = Number of Patients by Patient; Số Lượt Khám Theo Năm Sinh = Number of Visits by Year of Birth; Nhóm Bệnh có lượt khám cao theo giới tính = Number of patients by gender; Hướng Điều Trị = Treatment Direction; Phòng Khám Có Lượt Khám Cao Theo Giới Tính = Clinics with High Visits by Gender.

clouds, column charts, pie charts, or line charts (Figure 8). These visual tools clearly display commonly occurring diseases and their distribution. Word clouds visualize the frequency of diseases enabling users to identify disease trends and frequencies, facilitating appropriate prevention or treatment solutions to enhance healthcare quality and address specific medical challenges.

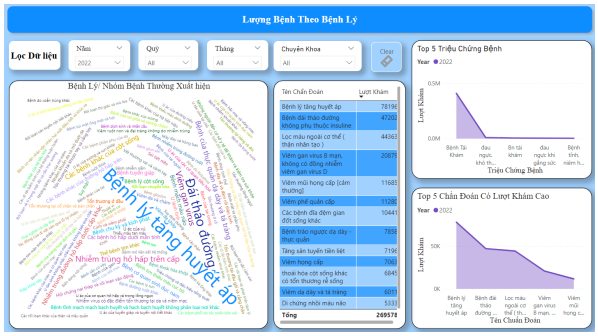

Fig. 8: Visualization of disease volume by category

Note: Lượng Bệnh Theo Bệnh Lý = Number of Patients by Disease; Bệnh Lý: Nhóm Bệnh Thường Xuất Hiện = Disease: Common Diseases; Top 5 Triệu Chứng Bệnh = Top 5 Symptoms; Top 5 Chẩn Đoán Có Lượt Khám Cao = Top 5 Diagnoses with Highest Visits.

Visualization of predicted clinical visit amount

Leveraging predictions from the previous section's LSTM model, the dashboard displays upcoming clinical visit volumes using line charts (Figure 9). Since planning decisions are often made on a monthly, quarterly, or bi-quarterly basis, monthly predictions were performed for the next six months. This intuitive visualization aids in management and planning, especially for resource allocation, staffing, and strategic decision-making to meet anticipated demand. Additionally, this feature predicts diseases or groups. This information supports management and planning, helps shape prevention strategies, and prepares for potential medical challenges. This feature is crucial for optimizing the health system's ability to predict and respond to expected disease trends.

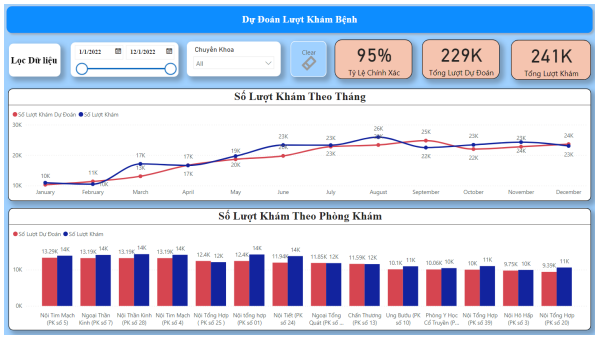

Fig. 9: Visualization of predicted clinical visit amount by month

Note: Dự Đoán Lượt Khám Bệnh = Predicted Number of Visits; Số Lượt Khám Theo Tháng = Number of Visits by Month; Số lượt Khám Dự Đoán = Predicted Visits; Số Lượt Khám = Actual Visits; Số Lượt Khám Theo Phòng Khám = Number of Visits by Clinic.

V. CONCLUSION

This study demonstrates the feasibility and usability of data analytics in healthcare management. By leveraging a data warehouse, critical insights provide actionable information for managers. The research findings indicate that LSTM models can effectively predict clinical visits, aiding in resource planning and operational efficiency. By predicting patient volume and disease trends. This model empowers clinics to: i) proactively allocate resources based on anticipated patient load; ii) develop effective plans to manage potential disease outbreaks; and iii) improve overall healthcare service delivery and patient care. To further enhance performance and applicability, future work should focus on upgrading predictive models with advanced deep learning techniques to improve accuracy. Integrating data from diverse sources will enrich the multidimensionality and depth of information, and expanding interface functionalities will improve user experience and utility. Ensuring the security and privacy of patient data through robust information security measures and strict

adherence to medical privacy regulations is also essential. Continued development is necessary to maintain integrity, accuracy, and efficiency, thereby fully supporting the management and analysis of clinical visit data.

REFERENCES

[1] Leung CK, Madill EWR, Tran NDT. Prediction of hospital status of COVID-19 patients from E-health records. In: *2022 IEEE International Conference on E-Health Networking, Application and Services, HealthCom 2022*. 17–19 October 2022; Genoa, Italy. IEEE; 2022. p.19–24. https://doi.org/10.1109/HealthCom54947.2022.9982738.

[2] Menon A, Aishwarya MS, Joykutty AM, Av AY, Av AY. Data visualization and predictive analysis for smart healthcare: Tool for a hospital. In: *2021 IEEE Region 10 Symposium (TENSYMP)*. 23–25 August 2021; Jeju, Republic of Korea. IEEE; 2021. p.1–8. https://doi.org/10.1109/TENSYMP52854.2021.9550822.

[3] Runkler TA. *Data analytics*. Springer; 2020.

[4] Zvonarev AE, Gudilin DS, Lychagin DA, Goryachkin BS. Extract-load-transform (ELT) process runtime analysis and optimization. In: *2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)*. 16–18 March 2023; Moscow, Russia. IEEE; 2023. p.1–7. https://doi.org/10.1109/REEPE57272.2023.10086728.

[5] Conn SS. OLTP and OLAP data integration: a review of feasible implementation methods and architectures for real time data analysis. In: *Proceedings IEEE SoutheastCon, 2005*. 08–10 April 2005; Ft. Lauderdale, FL, USA. IEEE; 2005. p.515–520. https://doi.org/10.1109/SECON.2005.1423297.

[6] Leung CK, Chen Y, Hoi CSH, Shang S, Cuzzocrea A. Machine learning and OLAP on big COVID-19 data. In: *2020 IEEE International Conference on Big Data (Big Data)*. 10–13 December 2020; Atlanta, GA, USA. IEEE; 2020. p.5118–5127. https://doi.org/10.1109/BigData50022.2020.9378407.

[7] Porwal P, Devare MH. Citation count prediction using different time series analysis models. In: *2022 IEEE Bombay Section Signature Conference (IBSSC)*. 08–10 December 2022; Bombay, India. IEEE; 2022. p.1–5. https://doi.org/10.1109/IBSSC56953.2022.10037553

[8] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*. 2019;31(7): 1235–1270. https://doi.org/10.1162/neco_a_01199.

[9] Yang S, Yu X, Zhou Y. LSTM and GRU neural network performance comparison study: taking Yelp review dataset as an example. In: *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*. 12–14 June 2020; Shanghai, China. IEEE; 2020. p.98–101. https://doi.org/10.1109/IWECAI50956.2020.00027.

[10] Luo J, Zhang Z, Fu Y, Rao F. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. *Results in Physics*. 2021;27: 104462. https://doi.org/https://doi.org/10.1016/j.rinp.2021.104462.

[11] Kapoor A, Sharma A. A comparison of short-term load forecasting techniques. In: *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*. 22–25 May 2018; Singapore. IEEE; 2018. p.1189–1194. https://doi.org/10.1109/ISGT-Asia.2018.8467788.

[12] Wang L, Wang H, Song Y, Wang Q. MCPL-based FT-LSTM: Medical representation learning-based clinical prediction model for time series events. IEEE Access. 2019;7: 70253–70264. https://doi.org/10.1109/ACCESS.2019.2919683.

[13] Men L, Ilk N, Tang X, Liu Y. Multi-disease prediction using LSTM recurrent neural networks. *Expert Systems with Applications*. 2021;177: 114905. https://doi.org/https://doi.org/10.1016/j.eswa.2021.114905.