

Research article

УДК 004.8

DOI: 10.17072/1993-0550-2026-1-92-99

<https://elibrary.ru/bexxas>



## IT Job Market Forecasting in East Africa: An ML Approach

Rebecca Ndungi<sup>1</sup>, Ivan Stanislavovich Blekanov<sup>2</sup>

<sup>1,2</sup> St. Petersburg State University, St. Petersburg, Russia

<sup>1</sup>[rebeccahndungi94@gmail.com](mailto:rebeccahndungi94@gmail.com)

<sup>2</sup>[i.blekanov@spbu.ru](mailto:i.blekanov@spbu.ru)

**Abstract.** This study focuses on forecasting the Information Technology (IT) job market in East Africa (specifically Kenya, Uganda, and Tanzania) using machine learning (ML) models. The research utilizes a dataset of 1,048,576 job postings collected from online platforms, including LinkedIn and Indeed. A comparative analysis of forecasting models Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Long Short-Term Memory (LSTM), and Holt's linear trend was conducted to predict employment trends, seasonality, and residual patterns. The models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The LSTM model demonstrated superior performance with an MAE of 2.75, MSE of 15.90, and RMSE of 3.99. The RMSE value of 3.99 indicates that the model's predictions are, on average, within approximately 4 job postings of the actual values. The findings confirm the applicability of ML models for reliable labor market forecasting in the region, providing valuable insights for stakeholders in education, policy, and industry to align strategies with market demands.

**Keywords:** *IT Industry; workforce strategy; job market analysis; technology demand forecasting; technological innovation demand forecasting; machine learning; Artificial Neural Network.*

**For citation:** Ndungi, R. and Blekanov, I. S. (2026), "IT Job Market Forecasting in East Africa: An ML Approach", *Bulletin of Perm University. Mathematics. Mechanics. Computer Science*, no 1(72), pp. 92–99, DOI: 10.17072/1993-0550-2026-1-92-99, <https://elibrary.ru/bexxas>.

*The article was submitted 09.04.2025; approved after reviewing 17.11.2025; accepted for publication 10.03.2026.*

Научная статья

## Прогнозирование рынка IT-вакансий в Восточной Африке: подход с использованием ML

Ндунги Р.<sup>1</sup>, Блеканов И. С.<sup>2</sup>

<sup>1,2</sup> Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

<sup>1</sup>[rebeccahndungi94@gmail.com](mailto:rebeccahndungi94@gmail.com)

<sup>2</sup>[i.blekanov@spbu.ru](mailto:i.blekanov@spbu.ru)

**Аннотация.** В данном исследовании разрабатывается подход к прогнозированию рынка вакансий в сфере информационных технологий (ИТ) в Восточной Африке (на



© Ndungi R., Blekanov I. S., 2026

Лицензировано по CC BY 4.0. Чтобы посмотреть копию этой лицензии, посетите <https://creativecommons.org/licenses/by/4.0/>

примере Кении, Уганды и Танзании) с применением моделей машинного обучения. В основе анализа лежит набор данных, содержащий 1 048 576 записей о вакансиях, собранных с онлайн-платформ, включая LinkedIn и Indeed. Проведено сравнительное исследование прогнозных моделей авторегрессионной интегрированной скользящей средней (ARIMA), сезонной ARIMA (SARIMA), сети с долгой краткосрочной памятью (LSTM) и линейного тренда Холта с целью прогнозирования динамики занятости, сезонных колебаний и остаточных паттернов. Для оценки эффективности моделей использовались метрики средней абсолютной ошибки (MAE), средней квадратичной ошибки (MSE) и среднеквадратической ошибки (RMSE). Наилучшие результаты показала модель LSTM, продемонстрировавшая значения MAE 2.75, MSE 15.90 и RMSE 3.99. Значение RMSE, равное 3.99, свидетельствует о том, что расхождения между прогнозируемыми и фактическими значениями в среднем составляют приблизительно 4 вакансии. Полученные результаты подтверждают применимость методов машинного обучения для достоверного прогнозирования регионального рынка труда и предоставляют стейкхолдерам в сфере образования, государственного управления и бизнеса ценную информацию для согласования стратегий с рыночными тенденциями.

**Ключевые слова:** *ИТ-отрасль; стратегия рабочей силы; анализ рынка труда; прогнозирование спроса на технологии и технологические инновации; машинное обучение; искусственная нейронная сеть.*

**Для цитирования:** Ндунги Р., Блеканов И. С. Прогнозирование рынка ИТ-вакансий в Восточной Африке: подход с использованием ML // Вестник Пермского университета. Математика. Механика. Информатика. 2026. № 1(72). С. 92–99. DOI: 10.17072/1993-0550-2026-1-92-99. <https://elibrary.ru/bexxas>.

*Статья поступила в редакцию 09.04.2025; одобрена после рецензирования 17.11.2025; принята к публикации 10.03.2026.*

## **Introduction**

The Information Technology (IT) sector represents a critical component of economic development in East Africa, characterized by rapid growth and changing skill requirements. Accurate forecasting of labor market dynamics within this sector has become essential for organizational strategy, educational planning, and policy formulation. While traditional forecasting techniques offer valuable insights, they often struggle to capture the complex, non-linear patterns that define contemporary job market data in emerging economies [1].

In time series forecasting, future values are estimated by the assessment of prior data, which has a timestamp. This approach is important in many fields because it helps ensure accurate decision-making and provide accurate forecasts. It involves using data gathered in the past to build models that will determine how an organization is going to chart its course in the future. Forecasting differs from other forms of analysis because producers cannot predict future events at the time of analysis. Instead, the process relies on assumptions that are derived from thorough research. Planning is enhanced by the use of statistical and modeling techniques to analyze data and make forecasts [1]. The uncertainty is inherent because of other factors, and larger datasets lead to greater accuracy. However, forecasts do offer an indication of the likelihood of specific occurrences [2].

This study focuses on an approach based on deep learning neural networks, intended to replace statistical methods such as the Auto-Regressive Integrated Moving Average (ARIMA) used in time series forecasting. The ARIMA model is recognized for its linearity and high accuracy in predicting results across various domains [3]. The model, an analytical instrument introduced by Box and Jenkins, is used more frequently in forecasting. However, there is a problem of decreased efficiency, especially when working with noisy data, this reducing the model's stability and accuracy [3]. The following three parameters need to be defined beforehand: the autonomy term ( $p$ ), the distinction term ( $d$ ), and the mean advance term ( $q$ ), where:

(p) stands for the number of previous values used in the formula, known as a lag order; (d) is the degree of difference used to transform the series to stationary; (q) reflects the size of the moving average smoothing or the order of the moving median [4].

A common approach to identifying the level of difficulty of a particular task is through the use of the forecasting horizon. While single-step forecasting is already quite difficult, multi-step forecasting presents even more difficulties such as the accumulation of errors, reduced precision, and enhanced variability [5]. In addition, it has its concept limitations on uncertainty and there may be consequences of unpredictability. While appropriate for some data, temporal series prediction is not appropriate across the board [5]. Since there is no consensus on the best forecasting methods, data teams and strategists need to be aware of the limitations and features of their computerized models. [5]. Data cleaning and timestamping are critical for accurate forecasting since they help one to determine the changes and trends in prior data. In addition to being able to parse fluctuations and discern trends from cyclical patterns, analysts need to detect outliers and fluctuations due to random events. It is believed that accurate predictions will shed light on the direction that these changes are likely to take in the future; looking at the progression of the time series analysis, it is possible to see a gradual evolution in people's perception of the world [5].

Machine learning presents a compelling alternative for time series forecasting, demonstrating a superior capacity to identify intricate, non-linear relationships that often challenge conventional statistical models. Among these advanced techniques, Long Short-Term Memory (LSTM) networks have gained prominence for sequence prediction due to their designed ability to capture and learn from long-range temporal dependencies. Machine learning algorithms are a source of several functions [6] such as sampling, classification, regression, and replication.

### **Research Problem and Methodological Gaps**

The East African IT job market presents unique forecasting challenges due to its dynamic nature and limited research coverage. Previous studies in labor market analysis exhibit several methodological limitations that our research addresses. Existing works often rely on single data sources, such as particular online job portals or Google Trends queries, which provide fragmented market perspectives and introduce source-specific bias [3, 8]. Many analyses remain predominantly retrospective, utilizing historical datasets that may not capture real-time market shifts driven by rapid technological adoption in the region. A significant research gap exists in a transparent rigorous comparison between modern machine learning approaches and established statistical benchmarks. While previous studies have applied advanced algorithms like Long Short-Term Memory (LSTM) networks, they often lack systematic evaluation against robust methods such as ARIMA and SARIMA for multi-step IT job forecasting [7, 9]. Furthermore, there is limited research focusing specifically on the East African context, where labor market dynamics differ substantially from developed economies. This study aims to address these gaps by developing a comprehensive forecasting framework that simultaneously ensures data comprehensiveness through multisource aggregation, enables transparent model comparison across different algorithmic families, and provides specific insights into the East African IT labor market.

### **Research Novelty and Contribution**

The primary objective of this research is to develop and validate a robust forecasting framework for the IT job market in East Africa through comparative analysis of statistical and machine learning models. The specific objectives include:

- acquisition and pre-processing of a large-scale, multi-source dataset of IT job postings from platforms including HeadHunter, LinkedIn, and regional job boards across Kenya, Uganda, and Tanzania;

- implementation of diverse forecasting approaches, including statistical methods (ARIMA, Seasonal ARIMA, Holt's linear trend) and deep learning models (LSTM);
- a rigorous comparative analysis using established error metrics (MAE, MSE, RMSE, MAPE, NRMSE) to objectively identify optimal forecasting approaches;
- extraction and analysis of the interpretable components trend, seasonality, and residuals within the East African IT job market dynamics.

The scientific novelty of this research lies in its integrated methodological approach that addresses multiple prior limitations simultaneously. The practical significance includes providing educational institutions with insights for curriculum development, enabling organizations to optimize talent acquisition strategies, and supporting policymakers in designing effective workforce development programs for the region's growing digital economy.

## **Methods and Materials**

### **Data collection and geographical scope**

This study utilizes a comprehensive dataset of 1,048,576 IT job postings collected from multiple online platforms between January 2020 and December 2023. Primary data sources include Indeed, LinkedIn, and regional job portals across East Africa, with specific focus on Kenya, Uganda, and Tanzania (East Africa). The dataset encompasses detailed job parameters including position titles, required skills, company information, location data, salary ranges, and posting dates. Data preprocessing followed the Pareto principle (80/20 rule) for feature selection and resource allocation [7]. The preprocessing pipeline included data cleaning, handling missing values, outlier management, timestamp normalization, and feature engineering to capture temporal patterns and regional characteristics specific to East African labor markets.

### **Forecasting methods**

#### **a. Long Short-Term Memory**

In the Train-Test Split phase, data is divided based on the sequence and rolling cross-validation. Model selection is the process of choosing the most suitable model when solving a particular problem. In model evaluation, different metrics, such as RMSE and MAE, are used to check the model's effectiveness. Where results contain confidence intervals. Model Deployment puts the generated model into practice in making real-time predictions, while monitoring and updating the model if necessary.

The modeling involved various algorithms such as LSTM, ARIMA, SARIMA, and Holt-Winter Multiplicative. This technique is based on the Autoregressive Integrated Moving Average (ARIMA), SARIMA, and the Exponential Smoothing models, which are very popular in analysis [7]. Since ARIMA models are particularly designed to capture temporal dependencies in data, they are suited to capture time-varying information such as job openings in a field like information technology. On the other hand, the complexity of Information Technology Business requires an enhanced approach based on improved methodology. The integration of a machine learning algorithm with the ARIMA model could act as a potent feature since the technology has the potential to learn from patterns. Neural networks endow the model with the ability to learn from experience and see patterns that are not discernable to a statistical model [5, 7].

#### **b. ARIMA and SARIMA**

These are two of the most commonly used models. One significant advantage of SARIMA is the ability to take seasonality into account in the received data. While SARIMA extends the basic ARIMA elements of the structure by adding seasonality and forming a decomposition of seasonally differenced values along with an autoregressive and moving average component, ARIMA essentially features only a simple form of the difference structure, as well as autoregressive and moving average components geared more towards non-seasonal data [9], [10]. This allows ARIMA to process co-variance better for the cases where seasonality changes over

time, making SARIMA invaluable for historical data that massively change during the year, for example, monthly retail sales or financial indicators for the quarters [8].

### **c. The Holt-Winter Multiplicative**

This method is one of the most popular forecasting techniques that is used in cases where trend and seasonality factors come into play categorized in two variants: additive and multiplicative variations. This forecasting method often goes by the name of the Holt-Winter technique and is a prevalent technique under the umbrella of exponential smoothing techniques. The reason for its popularity, however, lies in its simplicity, minimal space requirements, and the fact that its lender can be fully automated. However, in the Holt-Winter Multiplicative method, seasonality is handled as a relative factor or percentage rather than an absolute value, allowing seasonality to scale proportionally with the trend. Referring to LSTM as a type of RNN, the author notes that LSTM networks are "designed to overcome the difficulties of conventional RNNs with capturing long-term time dependencies" [11]. LSTMs use memory cells to store data for extended durations and control gates for data entry and output within the memory cells [12]. This architecture puts LSTMs in charge of processing repetitive tasks where temporal context is helpful [13]. For this reason, LSTMs are fitting for complex sequence prediction problems since these models and storage. Long Short-Term Memory (LSTM) is a popular artificial neural network technique that is being used a lot [11].

## **Results and Experiments**

### **Experiment description**

The model was trained using four different models, as shown in Table 1, with a dataset of 1,048,576 job records and was evaluated using different metrics including MAE, RMSE, NRMSE, MAE, and MAPE. Understanding patterns, regularity, and fluctuations in data is a crucial factor for making efficient decisions. The primary goal of forecasting when it comes to IT demand for jobs is to accurately predict the future market trends in employment based on previous information. By developing robust models that capture the essence of routine and cyclical activities, this research endeavors to provide reliable forecasts that inform consumers about strategic planning and informed decision-making. Accurate employment forecasting remains critical for ensuring that curricula, human capital development plans, and organizational hiring strategies in the IT field align with the required supply, thus minimizing imbalances in the job market.

The primary purpose of this research is to enhance the capability of the computing industry to improve both the responsiveness and the latency of forecasting functions within organizations. A key objective is to develop a system with the ability to approximate future values based on past data.

This involved the development of predictive models that capture the underlying characteristics of the data itself, thus producing accurate forecasts that are useful for decision-making. The construction of the predictive model involved the use of ARIMA, SARIMA, LSTM, and Holt-Winter Multiplicative methods. The dataset contains 1,048,576 jobs, and every job is presented through more than 15 fields. The proposed method was tested against three metrics, including MAE, MSE, and RMSE.

### **Experiment results**

The evaluation of the forecasting models has revealed distinct performance differences, quantified by standard error metrics. The results can be interpreted as follows: a lower value for all metrics (MAE, MSE, RMSE, MAPE, NRMSE) indicates a more accurate forecast; MAE (Mean Absolute Error) represents the average magnitude of the errors, in the same units as the original data, for instance, the volume of job postings; MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) penalize larger errors more severely. RMSE is also in the same units as the data, which makes it more interpretable than MSE.

MAPE (Mean Absolute Percentage Error) expresses the error as a percentage of the actual values, providing a scale-independent measure of accuracy. NRMSE (Normalized Root Mean Squared Error) allows for comparison between datasets with different scales. As shown in Table 1, the LSTM model demonstrated superior performance, achieving the lowest scores across all key metrics (MAE: 2.75, MSE: 15.90, RMSE: 3.99). This indicates that the LSTM predictions were, on average, closer to the actual values and contained fewer large errors compared to other models. For instance, its RMSE of 3.99 suggests its forecasts were typically within about 4 jobs of the actual value. While SARIMA was the best among the statistical models (MAE: 3.40, RMSE: 4.62), it was consistently outperformed by the LSTM. The traditional Holt's and ARIMA models showed the weakest results in this comparison.

**Table 1.** *Models used and the metric values*

<b>Model</b>	<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>	<b>MAPE</b>	<b>NRMSE</b>
ARIMA	3.80	24.60	4.96	0.0835	8.4498
SARIMA	3.40	21.30	4.62	0.8952	6.7381
LSTM	2.75	15.90	3.99	0.0762	5.4832
Holt's	3.60	23.80	4.88	0.9153	3.4523

**a. Holt's Additive Model**

Holt's Model includes the following measures: MAE was 3.60, MSE was 23.80, and RMSE was 4.88, as in Table 1 above. The values provide the estimated accuracy of the model and the magnitude of prediction errors.

**b. ARIMA Forecast**

The performance indicators for the ARIMA model using the formula below had a mean absolute error of 3.80, mean squared error of 24.60, and root of mean squared error of 4.96, as demonstrated in Table 1.

**c. SARIMA**

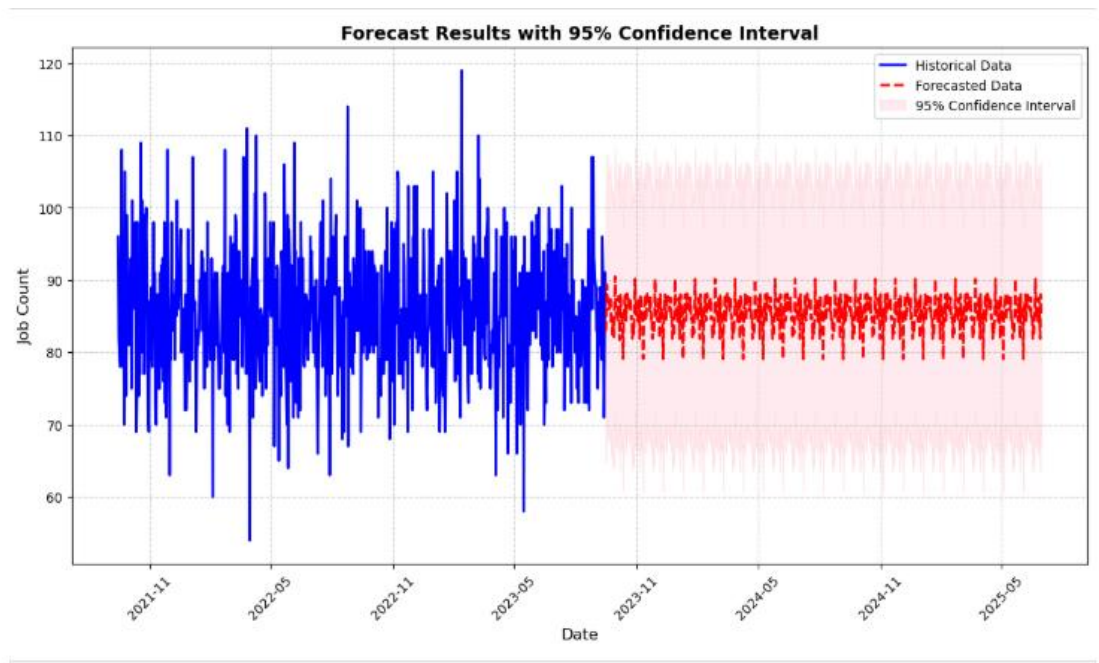
The SARIMA model had an MAE of 3.40, 21.30 MSE, and 4.62 RMSE. The metrics describe how accurate the model is in its predictions as well as the degree of error. The model shows improved values as compared to ARIMA, as observed in Table 1 above. Although the SARIMA model achieves reasonable accuracy, it is still possible to fine-tune its predictive capability for future fluctuations.

**d. LSTM Forecast with 100 Epochs**

Over 100 epochs, the LSTM method showed the training loss in gradual decline from 0.0320 before stabilizing at 0.0201. The training data set and the experimental data set had a smaller yield, which indicates the effectiveness of the training process with improved loss and reasonable generalization ability on unseen data. The model also yielded 2.75 MAE, 15.90 MSE and 3.99 RMSE, as seen in Table 1.

**Forecast Results**

The forecast for the next 6 months from the last observed date, as illustrated in Fig. 1, involved a study of the historical datasets on the trends and seasonality and shows the forecast area, the forecasted data, and the historical data section.



**Fig. 1.** Forecast results from the last observed data

## Conclusion

This research work establishes a comprehensive framework for IT job market forecasting in East Africa through a rigorous comparison of statistical and machine learning approaches. The demonstrated superiority of LSTM models in capturing complex market dynamics provides valuable methodology for labor market analysis in the region. The findings hold significant practical implications across multiple domains. Educational institutions can leverage these forecasting capabilities to align curriculum development with evolving regional industry demands. Organizations can benefit from enhanced workforce planning and talent acquisition strategies tailored to East African markets. Policymakers can gain valuable insights for regional development planning and educational policy formulation. Job seekers can make more informed decisions concerning skill development and career paths within the local context. Future research directions include implementation of sliding window techniques for enhanced model adaptability, development of hybrid modeling approaches combining statistical and machine learning methods, and extensive validation across other East African countries. Further investigations will explore real-time forecasting capabilities and integration of regional economic indicators for improved prediction accuracy to emerging economies.

## References

1. Senthurvelautham, S. and Senanayake, N. (2023), "A Machine Learning-Based Job Forecasting And Trend Analysis System To Predict Future Job Markets Using Historical Data", *IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, pp. 1–7, doi: 10.1109/I2CT57861.2023.10126233.
2. Li, W. and Law, K. L. E., (2024), "Deep Learning Models for Time Series Forecasting: A Review", *IEEE Access*, vol. 12, pp. 92306–92327, doi: 10.1109/ACCESS.2024.3422528.
3. Ray, S., Lama, A., Mishra, P., Biswas, T., Sankar Das, S. and Gurung, B. (2023), "An ARIMA-LSTM model for predicting volatile agricultural price series with random forest technique", *Appl Soft Comput*, vol. 49, p. 110939, doi: 10.1016/J.ASOC.2023.110939.
4. "Introduction to ARIMA models", (2024), URL: <https://people.duke.edu/~rnau/411arim.htm> (accessed: 02.04.2025).

5. Alrweili, H., Probab, H. F.-J. S. A. and undefined (2024), "Forecasting crude oil prices using an ARIMA-ANN hybrid model", *digitalcommons.aaru.edu.jo*, URL: <https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1505&context=jsap> (accessed: 02.04.2025).
6. Sarker, I. H., (2021), "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Comput Sci*, vol. 2, no 3, pp. 1–21, May 2021, doi: 10.1007/S42979-021-00592-X/FIGURES/11.
7. C. D.-I. J. of E. E. and Policy and undefined (2024), "The performance of hybrid ARIMA-GARCH modeling and forecasting oil price", *zbw.eu*, URL: <https://www.zbw.eu/econis-archiv/bitstream/11159/2094/1/1028123450.pdf> (accessed: 02.04.2025).
8. Lim, B. and Zohren, S. (2021), "Time-series forecasting with deep learning: A survey", *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences*, vol. 379, no 2194, doi: 10.1098/RSTA.2020.0209.
9. Hoong Lem, K. "The STL-ARIMA approach for seasonal time series forecast: a preliminary study", doi: 10.1051/itmconf/20246701008.
10. Tihi, N. and Popov, S. (2024), "The Higher Education Technical School of Professional Studies in Novi Sad, a comparison of Arima and random forest time series models for urban drought prediction", *Computer Science, and Data Science*, doi: 10.15308/Sinteza-2024-51-56.
11. Casolaro, A., Capone, V., Iannuzzo, G. and Camastra, F. (2023), "Deep Learning for Time Series Forecasting: Advances and Open Problems", *Information 2023*, Vol. 14, Page 598, vol. 14, no 11, p. 598, doi: 10.3390/INFO14110598.
12. Nazir, A. *et al.*, (2024), "A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem", *Ain Shams Engineering Journal*, vol. 15, no 7, p. 102777, doi: 10.1016/J.ASEJ.2024.102777.
13. Zhu, W. *et al.*, (2023), "An optimized long short-term memory (LSTM)-based approach applied to early warning and forecasting of ponding in the urban drainage system", *Hydrol Earth Syst Sci*, vol. 27, no 10, pp. 2035–2050, May 2023, doi: 10.5194/HESS-27-2035-2023.

**Information about the authors:**

R. Ndungi – Postgraduate student, St. Petersburg State University (7/9 University Embankment, St. Petersburg, Russia, 199034), ORCID: 0000-0002-0389-9758, Scopus Author ID: 58106191000, ResearcherID: RID66654;

I. S. Blekanov – Candidate of Technical Sciences; Associate Professor at St. Petersburg State University, Head of the Programming Technology Department, St. Petersburg State University (35/267 Universitetsky Prospekt, Peterhof, St. Petersburg, Russia, 198504), SPIN-code: 7473-1900, AuthorID: 700009, Scopus Author ID 56149559700, ORCID 0000-0002-7305-1429, ResearcherID G-8844-2015.

**Информация об авторах:**

Р. Ндунги – аспирант Санкт-Петербургского государственного университета (199034, Россия, г. Санкт-Петербург, Университетская набережная, д. 7/9), ORCID: 0000-0002-0389-9758, Scopus Author ID: 58106191000, ResearcherID: RID66654;

И. С. Блеканов – кандидат технических наук; доцент Санкт-Петербургского государственного университета, заведующий кафедрой технологии программирования (198504, Россия, г. Санкт-Петербург, Петергоф, Университетский проспект, д. 35, ком. 267), SPIN-код: 7473-1900, AuthorID: 700009, Scopus Author ID 56149559700, ORCID 0000-0002-7305-1429, ResearcherID G-8844-2015.