Forecasting the Risk of Cervical Cancer in Women in the Human Capital Development Context Using Machine Learning

Iryna Bodnar[1], Myroslava Bublyk[2[0000-0003-2403-0784]], Oleh Veres[3[0000-0001-9149-4752]], Olga Lozynska[4[0000-0002-5079-0544]], Ihor Karpov[5[0000-0003-4885-5078]], Yevhen Burov[6[0000-0001-6124-3995]], Petro Kravets[7[0000-0001-6124-3995]], Ivan Peleshchak[8[0000-0002-7481-8628]], Olena Vovk[9[0000-0001-5523-0901]], and Olga Maslak[10[0000-0001-6793-4367]]

[1-9]Lviv Polytechnic National University, Lviv, Ukraine
[10]Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine
iryna.bodnar.sa.2017@lpnu.ua[1], my.bublyk@gmail.com[2],
olena.b.vovk@lpnu.ua[8], oimaslak2017@gmail.com[9]

Abstract. This work offers a model for predicting the risk of a particular disease in a particular patient. This work is aimed at improving practical skills in implementing machine-learning methods, as well as creating a model that would be based on an appropriate ML algorithm and give a clear forecast of the risk of cancer in women.

Keywords: Cervical Cancer, Forecasting the Risk, Machine Learning, Human Capital.

1   Introduction

The 21 century is characterized by accelerated international integration processes of countries, where human capital plays a key role in the global space. The 2030 Agenda for Sustainable Development of the UN General Assembly adopted the Resolution [1] determining key directions of development. The document lays out the 17 Sustainable Development Goals, among which is Goal #3 entitled "Good health and well-being". However, lapses in environmental health are increasingly recognized as major contributors to the illness and death of people, therefore, the problem of forecasting the risk of cervical cancer in women is actually today. The problem of morbidity and mortality from tumors, according to [2], is directly related to establishing the pollution impact on the mortality rate from tumors and predicting the risks of this disease.

However, cervical cancer is caused by a long-term infection in the cells of the skin and mucous membrane of the genitals, namely the human papillomavirus [3-4], therefore, it is one of the most deadly diseases in the world among women. The danger further of this cancer is that it does not show any symptoms, in the early stages of the disease [5].

The project is created to facilitate the work of doctors and simplify the identification of those patients who are in the highest risk zone. This work is based on a data set

consisting of both input and target variables based on the anamnesis of many patients [3-4]. The best subset of the data set among all subsets of objects in experiments was investigated. The random forest method was used to predict the risk of disease occurrence [5-9].

2 Related Work

Cancer is the leading cause of death worldwide; it causes 8.8 million deaths in 2019. Cancer occurs when normal cells turn into tumor cells in a multi-stage process that usually progresses from a cancerous lesion to a malignant tumor. With early detection, cancer responds faster to effective treatment, which can increase the likelihood of survival, reduce the incidence, and treatment costs. Significant improvements in a patient's life can be achieved by early detection of cancer and avoiding delays in care [3, 13-20]. To reduce unnecessary examinations, patients can take a survey to determine if they belong to the risk group. Therefore, based on this study, it will be possible to create a plan for examining patients, according to a certain risk.

3 Brief Description of the Data Used

The dataset was obtained from the Machine Learning Repository dataset archive (Fig. 1). The data set was collected at the Universitario de Caracas hospital in Caracas, Venezuela. The data set includes demographic information, habits, and medical records of 858 patients. There are 35 attributes in the data set. The attributes of the data set are shown in Fig. 1 [4].

Fig. 1. Displaying a data set

4 Choice of Research Methods and Tools

The method used in this work is the Random forest. Random forest is an ensemble method of machine learning for classification, regression and other tasks that operate by building multiple decision trees during model training and produce a fashion for classes (classifications) or averaged forecast (regression) of constructed trees [21-32]. Object classification is done by voting: each Committee tree assigns an object that is classified to one of the classes and the class that the largest number of trees voted for wins. The authors of [21-32] consider the tendency to retraining to be a disadvantage of the Random forest method.

The Python programming language is a high-level object-oriented language that allows you to use classes and their objects to build the structure of the software you are developing. This language is interpreted, not compiled, which allows you to execute code "on the fly" [25]. In addition, Python is best suited for performing machine learning tasks, because it is quite clear compared to other languages. Moreover, it has the excellent data processing performance [6, 25]. One of the main reasons why Python is used for machine learning is that it has many frameworks that simplify the process of writing code and reduce development time.

The implementation of the software took place with the help of the program Microsoft Visual Studio. Visual Studio is an integrated software development environment from Microsoft. This environment allows you to create a variety of software products: console programs, programs with a graphical interface, such as Windows Forms window applications, as well as Web applications, and so on [33].

To simplify the work, we used the following built-in Python libraries [34-41]:

- Numpy is an extension of the Python language that supports large multidimensional arrays and matrices, along with a large library of high-level mathematical functions for operations with these arrays.
- Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating multiple tables and time series.
- Matplotlib is a Python programming language library for visualizing data with two-dimensional 2D graphics (3D graphics are also supported). The resulting images can be used as illustrations in publications.
- Math is one of the most important in Python. This module provides a lot of functionality for working with numbers.
- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- The Scikit-learn library is the most common choice for solving classical machine learning problems. It provides a wide range of learning algorithms with and without a teacher.

5   Description of the Control Example

The first step before building a model is to analyze the available data. This action is necessary in order to establish certain interdependencies between the specified characteristics. The uploaded data (Fig. 2) and its main characteristics (Fig. 3) are shown below.

Fig. 2. The downloaded data

Fig. 3. Information about each attribute

After analyzing all the graphs of the distribution of attributes, it is clear that the column of hormonal contraceptives has the highest number of positive results, which indicates that this may be an important key feature of detecting cervical cancer. The distribution hormonal contraceptives is shown in Fig.4
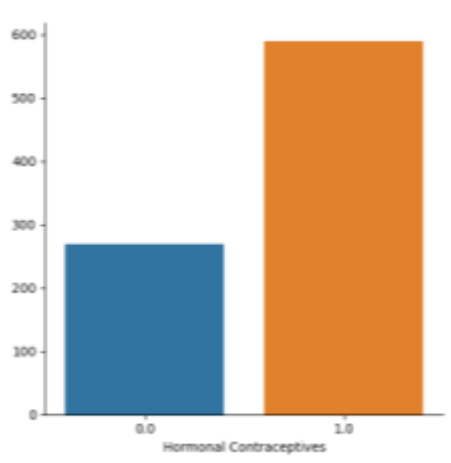
Fig. 4. Distribution of hormonal contraceptives

The analysis of the age attribute was conducted to determine the age of women most often facing this problem. Fig. 5 shows the distribution of the patient number versus their age. The average age of patients with a confirmed diagnosis of cervical cancer was found to be 26.8 years.
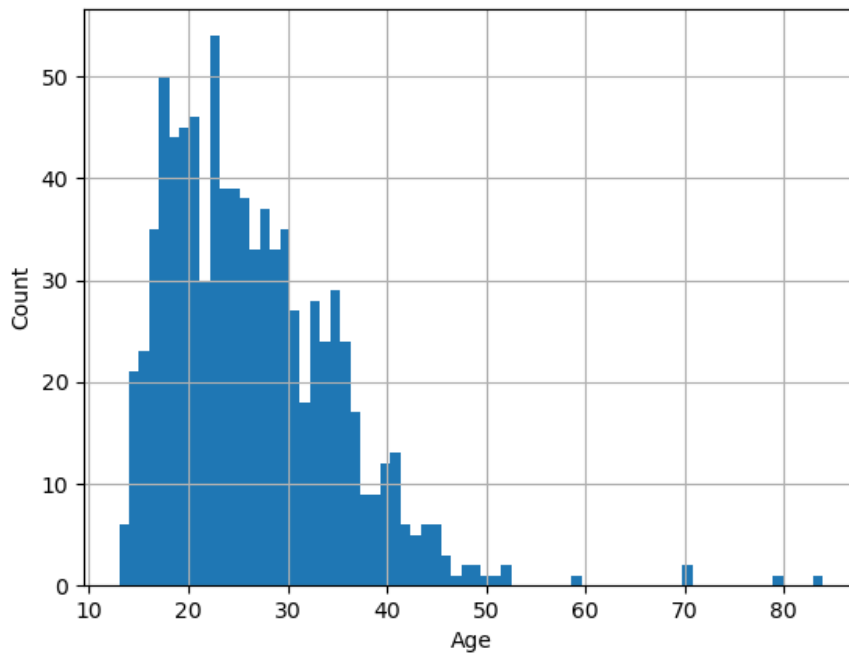
Fig. 5. Distribution of the patient number versus their age

6 Analysis of the Obtained Results

Density plots (Fig. 6-9) show that women who started sexual activity at the age of 15 to 20 years have a high chance of developing a malignant tumor due to incomplete development of the corresponding organs. Therefore, this characteristic can also be considered as a key feature in the detection of cervical cancer.
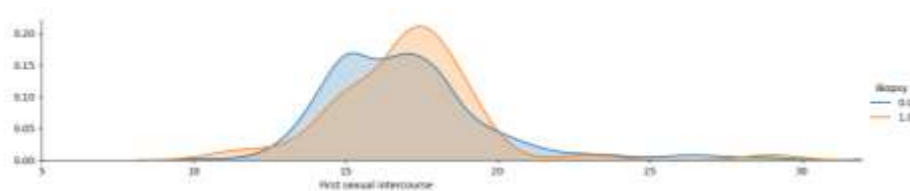
Fig. 6. Density Plots for Biopsy tests and First sexual intercourse attributes
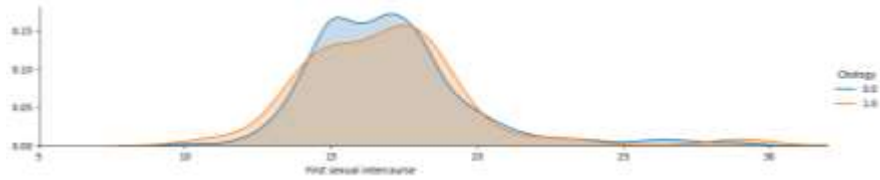
Fig. 7. Density Plots for Cytology tests and First sexual intercourse attributes
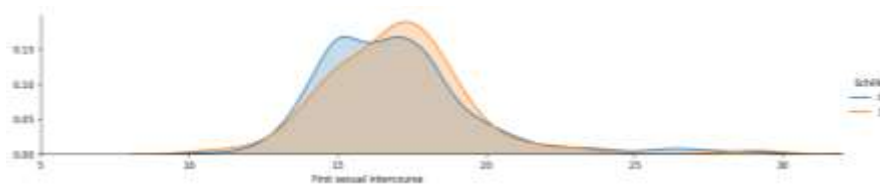
Fig. 8. Density Plots for Schiller tests and First sexual intercourse attributes
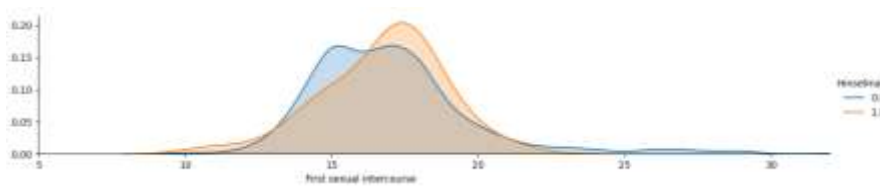
Fig. 9. Density Plots for Hinselmann tests and First sexual intercourse attributes

The conducted analysis has shown that of the three columns (Smokes, Smokes (years), and Smokes (packs/ year), the smokes column can be deleted. This is due Smokes (years) will represent the Smokes column (since the Smoke (years) column has a null value only when the Smokes column has a null value). Also, forecasts depend on the number of years of smoking, and not on whether a person did it or not. Based on the distribution of data points at the scatter site, the Schiller and Biopsy tests show that women who smoke for more than 5 years are more likely to develop a malignant tumor (Fig. 10-11).
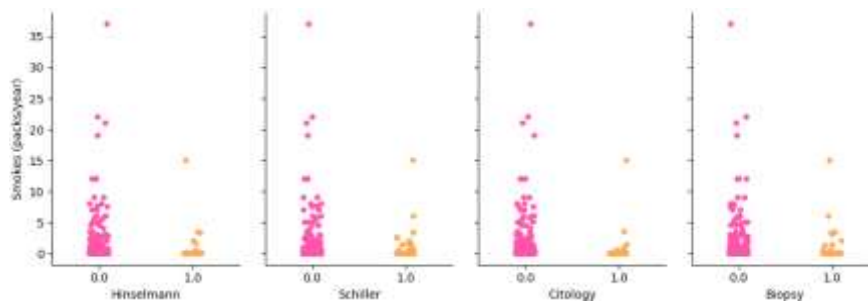
Fig. 10. Scatter Plots of target variables depend on Smokes (packs/years)
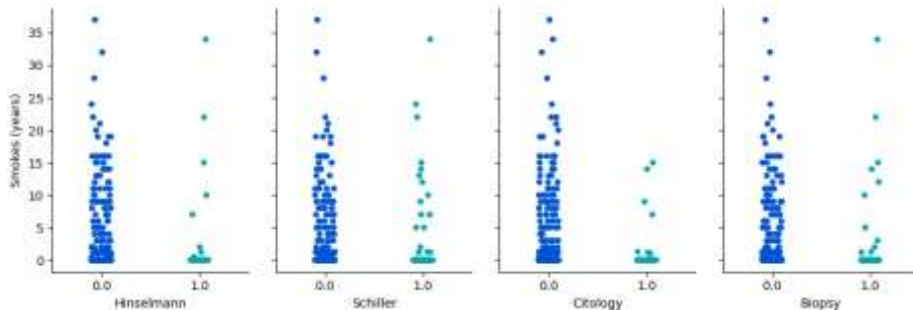
Fig. 11. Scatter Plots of target variables depend on Smokes (years)

The distribution of negative test results shows that an intrauterine device (IUD) used for any number of years has an effect on reducing the chances of encountering cancer (Fig. 12).
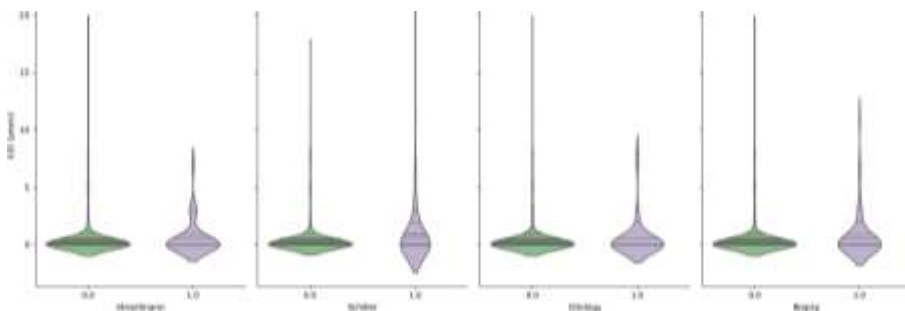
Fig. 12. The distribution of target variables with respect to IUD (years)

After a detailed analysis of the characteristics, we moved on to building a model. There are 4 target variables in the dataset that can take the values of 0 or 1. To create the model, we entered a new variable that gives five different values, ranging from 0 to 4. These values represent the risk level, meaning 0 for indicating no risk, and 4 indicating the highest level of risk. So, the final values for the new variable can take 0, 1, 2, 3, 4. After that, we counted the number of each value (Fig. 13) and concluded that the data available is not enough.

```
Value counts of each target variable: 0.0
1.0     41
3.0     33
2.0     22
4.0      6
Name: cervical_cancer, dtype: int64
Final feature vector shape: (858, 26)
Final target vector shape (858,)
```

Fig. 13. Value counts of each target variable

So the Smote () method was used, which is based on the idea of creating a certain number of artificial examples that would be similar to those in the sample, but would not duplicate them. Next, we calculated the accuracy of models created by various algorithms (logistic regression with the parameter multi_class='ovr' and 'multinomial', the method of support vector machines with the classifiers SVC and LinearSVC and Random forest) using the cross-checking method (Fig.14). It aims to check whether the results of statistical analysis are generalized on are an independent data set).

```
Results of the cross validation on dataset with 26 features:

LogisticRegMulti: 0.531 (0.020)
LogisticRegOVR: 0.498 (0.026)
svm_linear: 0.390 (0.099)
svm_rbf: 0.837 (0.018)
RandomForest: 0.944 (0.031)
```

Fig. 14. The calculated accuracy for each algorithm

From the results obtained in the previous step, we chose the most correct algorithm for the data, namely Random Forest, whose accuracy is 0.944, and predicted the threat level for three women with different indicators. The results are shown in Fig. 15.

```
Testing the model on women with age less than 20: [0]
Testing the model on women with age 35: [4]
Testing the model on women with age 48: [1]
Press any key to continue . . .
```

Fig. 15. The predicted risk of disease for 3 women

- Women in the age of 19, Number of sexual partners: 1, First sexual intercourse: 17, Num of pregnancies: 1, Smokes (years): 1, Smokes (packs/year): 3.4, Hormonal Contraceptives (years): 0 and all the other features are 0;
- Women with an age of 35, Number of sexual partners: 5,First sexual intercourse: 11, Num of pregnancies: 2, Smokes (years): 15, Smokes (packs/year): 15, Hormonal Contraceptives (years): 0 and all the other features are 0;
- Women with an age of 48, Number of sexual partners: 2, First sexual intercourse: 15, Num of pregnancies: 2, Smokes (years): 0, Smokes (packs/year): 0, Hormonal Contraceptives (years): 0.5, IUD (years): 19, STDs: genital herpes:1 and all the other features are 0.

7   Conclusions

This article concerns prediction of the risk of cervical cancer. The main provisions of the study were reviewed and briefly described. The data set used for forecasting is also described in detail. After that, the method and means of implementation are described. A method for creating was a Random Forest machine learning model. The implementation tool was the Python programming language, because it is frequently used for machine learning, due to the wide range of its frameworks, what simplifies the process of writing code and reduce development time. The Microsoft Visual Studio software made the implementation of the software.

At the end of this article, a control example is analyzed to determine the accuracy of the created model. The value of the cross validation coefficient for the created model is 0.944, what means that the results obtained are highly accurate.

References

1. The 2030 Agenda for Sustainable Development: Transforming our World: the 2030 Agenda for Sustainable Development, United Nations, New York. (2019)
2. Bublyk, M.I., Rybytska, O. M.: The model of fuzzy expert system for establishing the pollution impact on the mortality rate in Ukraine. In: Computer sciences and information technologies: Proceedings of the 2017 12th International Scientific and Technical Conference (CSIT 2017), 1, 253–256. (2017)
3. Jaswinder Singh, Sandeep Sharma: Determining Cervical Cancer Possibility by Using Machine Learning Methods, https://www.researchgate.net/publication/322233711_Determining_Cervical_Cancer_Possibility_by_Using_Machine_Learning_Methods
4. Cervical cancer (Risk Factors) Data Set, https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
5. Kuzmin, O., Bublyk, M.: Economic evaluation and government regulation of technogenic (man-made) damage in the national economy. In: Computer sciences and information technologies. Proceedings of the XIth International Scientific and Technical Conference (CSIT 2016), 37–39. (2016). DOI: https://doi.org/10.1109/STC-CSIT.2016.7589863
6. Python for Machine Learning, https://medium.com/nuances-of-programming/%D0%BF%D0%BE%D1%87%D0%B5%D0%BC%D1%83-python-%D0%B8%D1%81%D0%BF%D0%BE%D0%BB%D1%8C%D0%B7%D1%83%D0%B5%D1%82%D1%81%D1%8F-%D0%B4%D0%BB%D1%8F-%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE-%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%8F-bc1a2d8f5963
7. Chyrun, L., Leshchynskyy, E., Lytvyn, V., Rzheuskyi, A.,Vysotska, V., Borzov, Y.: Intellectual Analysis of Making Decisions Tree in Information Systems of Screening Observation for Immunological Patients. In: CEUR Workshop Proceedings, Vol-2362, 281-296. (2019)
8. Lytvyn, V., Burov, Y., Kravets, P., Vysotska, V., Demchuk, A., Berko, A., Ryshkovets, Y., Shcherbak, S., Naum, O.: Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification. In: CEUR Workshop Proceedings, Vol-2362, 354-368. (2019)

9. Vysotska, V., Lytvyn, V., Burov, Y., Gozhyj, A., Makara, S.: The consolidated information web-resource about pharmacy networks in city. In: CEUR Workshop Proceedings, 239-255. (2018)
10. Cherednichenko, O., Kanishcheva, O., Yakovleva, O., Arkatov, D.: Collection and Processing of a Medical Corpus in Ukrainian. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 272-282. (2020)
11. Perkhach, R., Kysil, D., Dosyn, D., Zavuschak, I., Kis Y., Hrendus, M., Vasyliuk, A., Sadova, M., Prodanyuk, M.: Method of Structural Semantic Analysis of Dental Terms in the Instructions for Medical Preparations. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 662-669. (2020)
12. Cherednichenko, O., Babkova, N., Kanishcheva, O.: Complex Term Identification for Ukrainian Medical Texts. In: CEUR Workshop Proceedings, Vol-2255, 146-154. (2018)
13. Fedushko, S.: Adequacy of Personal Medical Profiles Data in Medical Information Decision-Making Support System. In: CEUR Workshop Proceedings, Vol-2544. (2020)
14. Fedushko, S., Gregus, Michal ml.: Ustyianovych T. Medical card data imputation and patient psychological and behavioral profile construction. In: The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH), 160, 354-361. (2019)
15. Fedushko, S., Syerova, Yu.: Classification of Medical Online Helpdesk Users. In: CEUR Workshop Proceedings, Vol 2544. (2020)
16. Shakhovska, N., Fedushko, S., Greguš, ml. M., Melnykova, N., Shvorob, I., Syerov, Yu.: Big Data analysis in development of personalized medical system. In: The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN), 160, 229-234. (2019)
17. Shakhovska, N., Fedushko, S., Greguš, ml. M., Shvorob, I., Syerova, Yu.: Development of Mobile System for Medical Recommendations. In: The 15th International Conference on Mobile Systems and Pervasive Computing (MobiSPC), 155, 43-50. (2019)
18. Shakhovska, N., Shakhovska, K., Fedushko, S.: Some Aspects of the Method for Tourist Route Creation. In: Advances in Artificial Systems for Medicine and Education II, 902, 527-537. (2019)
19. Syerov, Y., Shakhovska, N., Fedushko, S.: Method of the Data Adequacy Determination of Personal Medical Profiles. In: Advances in Artificial Systems for Medicine and Education II, 902, 333-343. (2019)
20. Antonyuk N., Chyrun L., Andrunyk V., Vasevych A., Chyrun S., Gozhyj A., Kalinina I., Borzov Y.: Medical News Aggregation and Ranking of Taking into Account the User Needs. In: CEUR Workshop Proceedings, Vol-2362, 369-382. (2019)
21. Burov, Y., Vysotska, V., Kravets, P. Ontological approach to plot analysis and modeling. In: CEUR Workshop Proceedings, Vol-2362, 22-31. (2019)
22. Vysotska, V., Lytvyn, V., Burov, Y., Berezin, P., Emmerich, M., Fernandes, V. B.: Development of Information System for Textual Content Categorizing Based on Ontology. In: CEUR Workshop Proceedings, Vol-2362, 53-70. (2019)
23. Lytvyn, V., Vysotska, V., Rusyn, B., Pohreliuk, L., Berezin, P., Naum O.: Textual Content Categorizing Technology Development Based on Ontology. In: CEUR Workshop Proceedings, Vol-2386, 234-254. (2019)
24. Kravets, P., Lytvyn, V., Vysotska, V., Burov, Y.: Promoting training of multi-agent systems. In: CEUR Workshop Proceedings, Vol-2608, 364-378. (2020)
25. Matseliukh, Y., Vysotska, V., Bublyk, M.: Intelligent system of visual simulation of passenger flows. In: CEUR Workshop Proceedings, 2604, 906. (2020)

26. Bublyk, M., Matseliukh, Y., Motorniuk, U., Terebukh, M.: Intelligent system of passenger transportation by autopiloted electric buses in Smart City. In: CEUR Workshop Proceedings, 2604, 1280. (2020) http://ceur-ws.org/Vol-2604/paper81.pdf

27. Lytvyn, V., Vysotska, V., Osypov, M., Slyusarchuk, O., Slyusarchuk, Y.: Development of intellectual system for data de-duplication and distribution in cloud storage. In: Webology, 16(2), pp. 1-42. (2019)

28. Kravets, P., Burov, Y., Lytvyn, V., Vysotska, V.: Gaming method of ontology clusterization. In: Webology, 16(1), 55-76. (2019)

29. Emmerich, M., Lytvyn, V., Yevseyeva, I., Fernandes, V. B., Dosyn, D., Vysotska, V.: Preface: Modern Machine Learning Technologies and Data Science (MoMLeT&DS-2019). In: CEUR Workshop Proceedings, Vol-2386. (2019)

30. Melnykova, N., Markiv, O.: Semantic approach to personalization of medical data. In: Computer Sciences and Information Technologies, CSIT, 59-61. (2016)

31. Melnykova, N., Shakhovska, N., Sviridova, T.: The personalized approach in a medical decentralized diagnostic and treatment. In: 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics, 295-297. (2017)

32. Kryvenchuk, Y., Shakhovska, N., Shvorob, I., Montenegro, S., Nechepurenko, M.: The smart house based system for the collection and analysis of medical data. In: CEUR Workshop Proceedings, 2255, 205-214. (2018)

33. Visual Studio documentation, https://docs.microsoft.com/en-us/visualstudio/windows/?view=vs-2019

34. Sarkar, D., Bali, R., Sharma, T.: Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems, Apress, Berkeley. (2018)

35. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer. (2018), http://www.springer.com/in/book/9781461471370

36. Boyko, N., Sviridova, T., Shakhovska, N.: Use of machine learning in the forecast of clinical consequences of cancer diseases. In: Mediterranean Conference on Embedded Computing, MECO - Including ECYPS, 1-6. (2018)

37. Malachivskyy, P.S., Pizyur, Y.V., Andrunyk, V.A.: Chebyshev Approximation by the Sum of the Polynomial and Logarithmic Expression with Hermite Interpolation. In: Cybernetics and Systems Analysis, 54(5), 765-770. (2018)

38. Veres, O., Rishnyak, I., Rishniak, H.: Application of Methods of Machine Learning for the Recognition of Mathematical Expressions. In: CEUR Workshop Proceedings, Vol-2362, 378-389. (2019)

39. Andrunyk, V., Berko, A., Rusyn, B., Pohreliuk, L., Chyrun, S., Dokhniak, B., Karpov, I., Krylyshyn, M.: Information System of Photostock Web Galleries Based on Machine Learning Technology. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 1032-1059. (2020).

40. Zdebskyi, P., Lytvyn, V., Burov, Y., Rybchak, Z., Kravets, P., Lozynska, O., Holoshchuk, R., Kubinska, S., Dmytriv, A.: Intelligent System for Semantically Similar Sentences Identification and Generation Based on Machine Learning Methods. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 317-346. (2020).

41. Makara, S., Chyrun, L., Burov, Y., Rybchak, Z., Peleshchak, I., Peleshchak, R., Holoshchuk, R., Kubinska, S., Dmytriv, A.: An Intelligent System for Generating End-User Symptom Recommendations Based on Machine Learning Technology. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 844-883. (2020).