DEVELOPMENT OF METHODS AND TOOLS FOR ANOMALY DETECTION IN DATA ANALYSIS TASKS

Dmytro Palahin

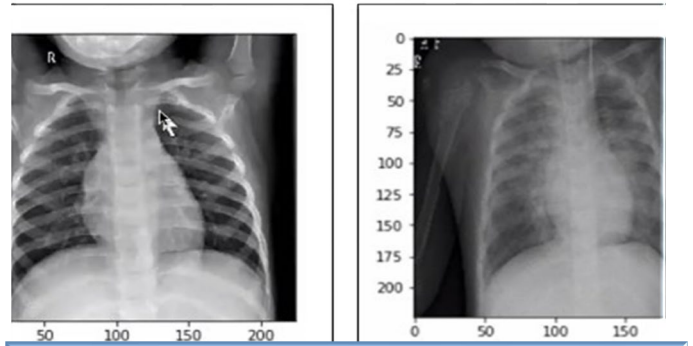
Object and Subject of Research

Machine learning methods for anomaly detection in data, software tools for algorithmic implementation of anomaly detection processes

Why am I interested in this topic?

Topicality

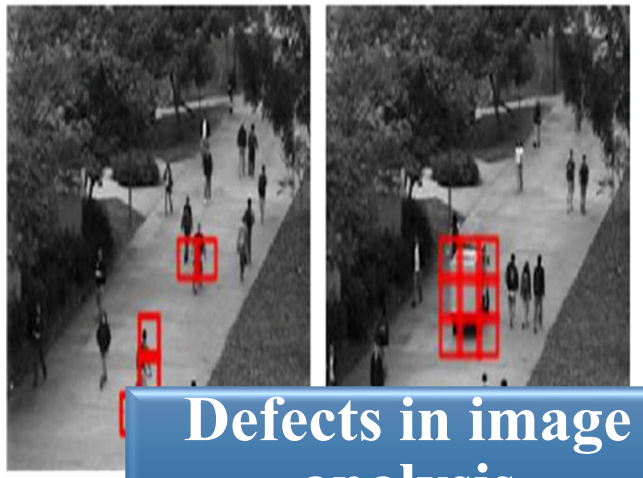
Fraud detection



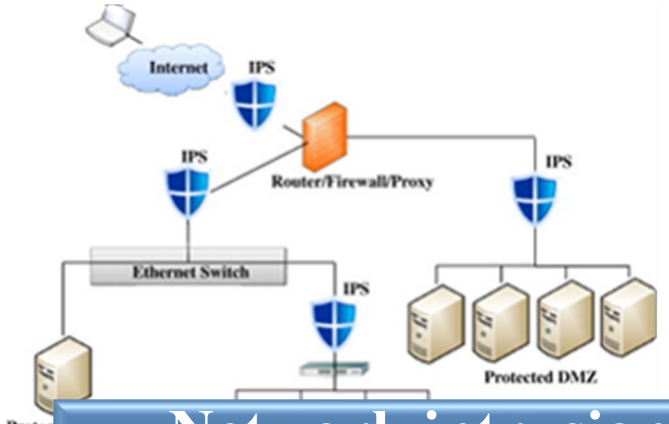
Medical diagnosis

Detection of production defects

Online shopping

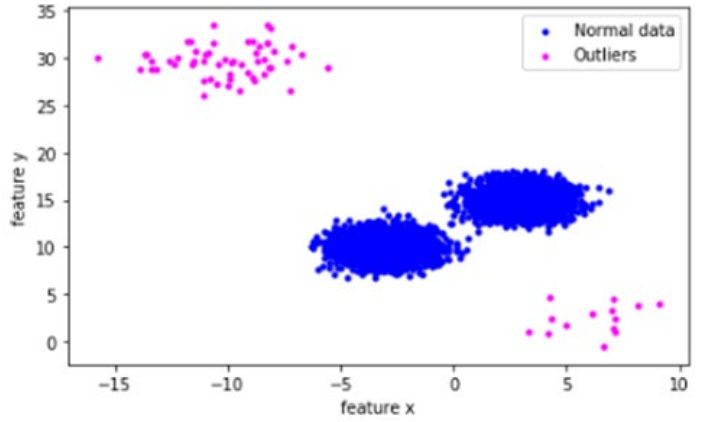


Defects in image analysis

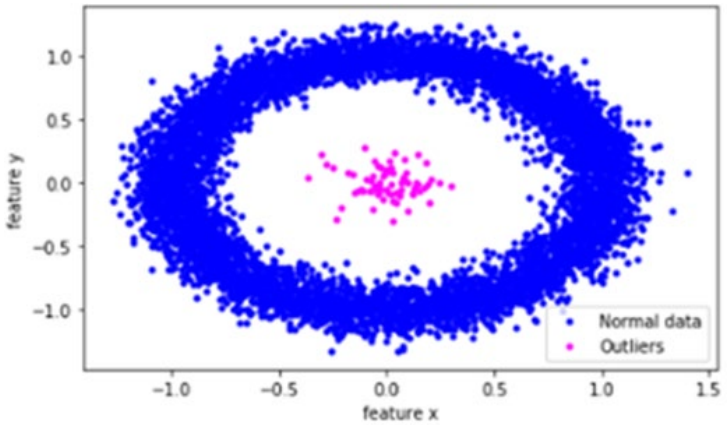


Network intrusion detection

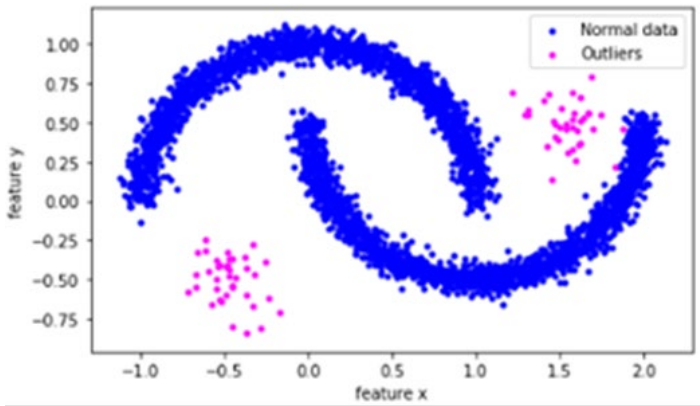
Model Data Structures



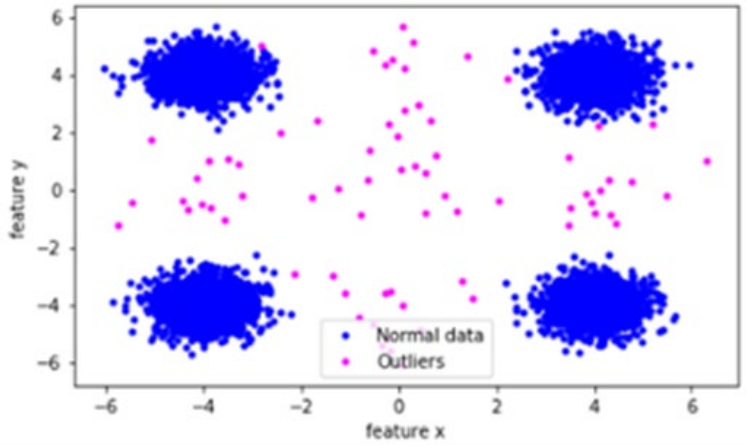
2 Blobs



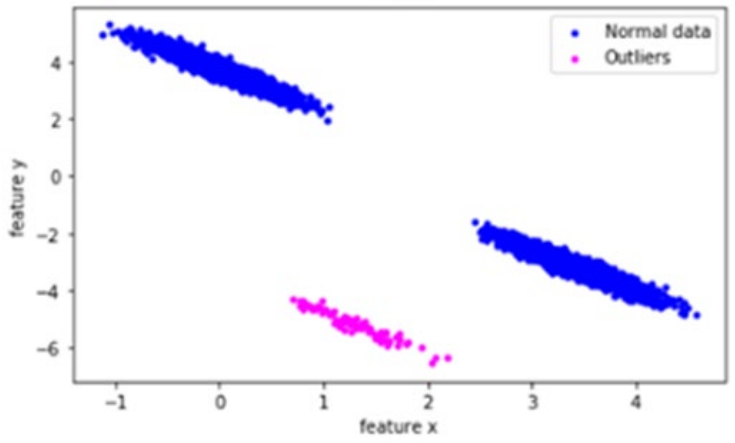
Circles



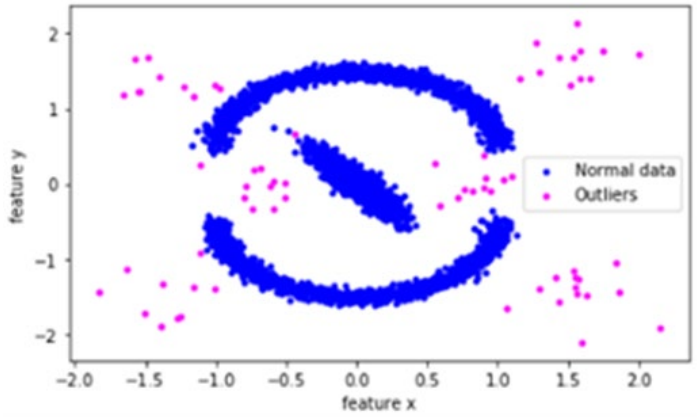
Moons



4 Blobs

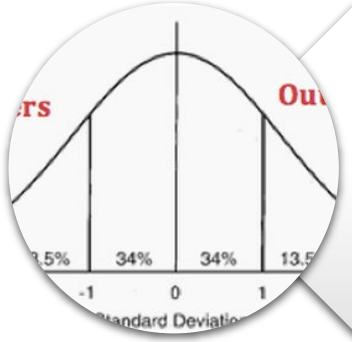


Aniso



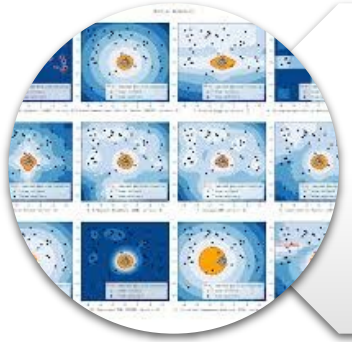
Moons +

Methods of Detecting Anomalies in Data



Statistical methods:
- 3-Sigma;
- Interquartile range;
- Z-score;
- Mahalanobis distance.

Methods of the Scikit-Learn library:
- DBSCAN;
- IsolationForest;
- LocalOutlierFactor;
- OneClassSVM and other.

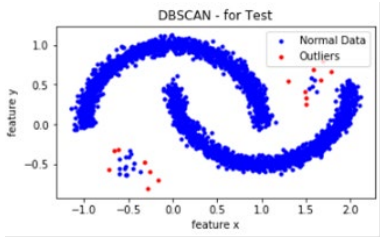


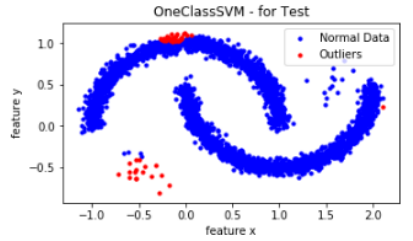
Methods of the PyOD library :
- CBLOF;
- XGBOD;
- AutoEncoder;
- KNN and other.

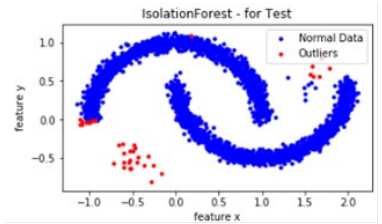
The Result of Processing the Data Structure of the "Moons" Type

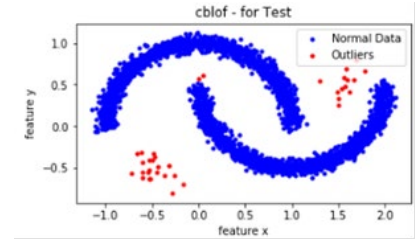
Indicators of the effectiveness of anomaly detection of the "MOONS" type data structure by various methods

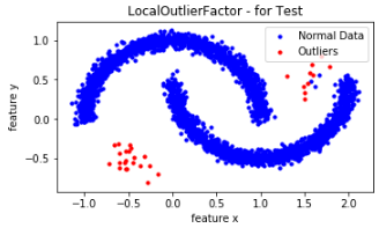
№	Method name	Performance indicator				
		f1	**roc_auc**	**accuracy**	**precision**	**recall**
Statistical methods	**3 Sigma**	0.497	0.5	0.988	0.494	0.5
	Quantile	0.497	0.5	0.988	0.494	0.5
	Z-score	0.497	0.5	0.988	0.494	0.5
	Mahalanobis distance	0.735	0.817	0.984	0.687	0.817
Scikit-Learn library	***DBSCAN***	0.804	0.720	0.993	0.996	0.720
	IsolationForest	0.873	0.895	0.994	0.854	0.895
	LocalOutlierFactor	0.976	0.955	0.999	0.999	0.955
	OneClassSVM	0.698	0.744	0.983	0.667	0.744
PyOD library	***CBLOF***	0.985	0.999	0.999	0.972	0.999
	XGBOD	0.992	0.999	0.999	0.985	0.999
	AutoEncoder	0.743	0.732	0.989	0.755	0.732
	KNN	0.935	0.998	0.996	0.886	0.998

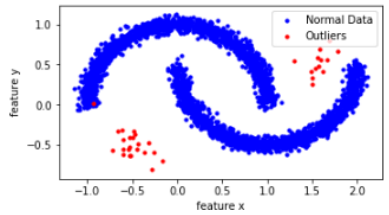


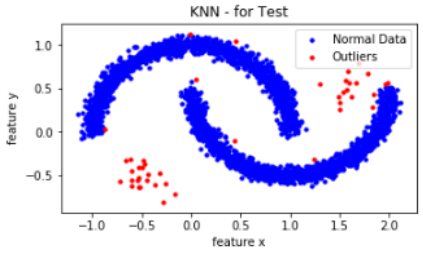


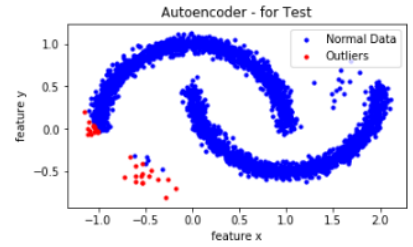






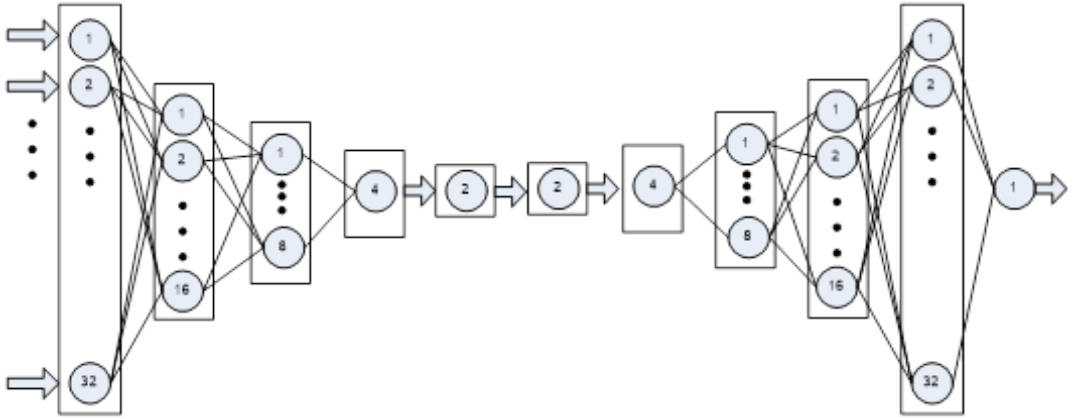




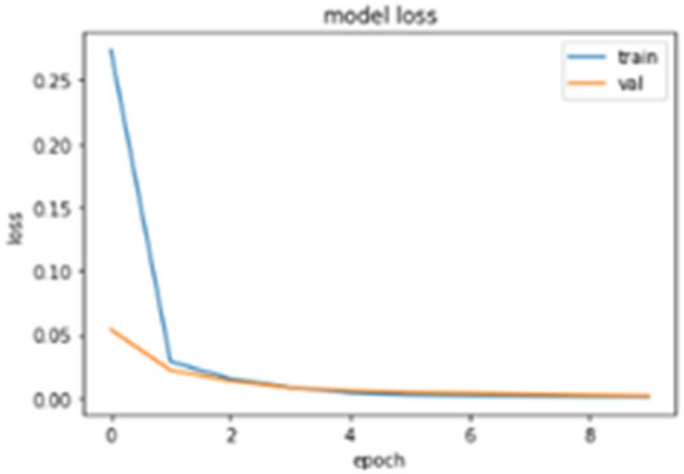


But these methods do not always show an effective result, in particular AutoEncoder from the **PyOD library**

I. Development of the Autoencoder Classifier Neural Network Structure



Dense Layer Autoencoder Model – "Autoencoder Classifier"



Learning Process of the Neural Network - "Autoencoder Classifier"

Indicators of anomaly detection efficiency for various data structures when applying a new neural network structure

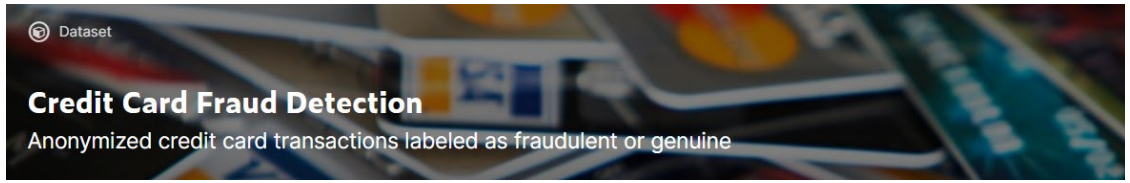
Data structure name	Performance indicator				
	f1	**roc_auc**	**accuracy**	**precision**	**recall**
2 BLOBS	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
CIRCLES	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
MOONS	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
4 BLOBS	0.985	0.971	0.999	0.999	0.971
ANISO	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
MOONS +	0.884	0.814	0.995	0.997	0.814

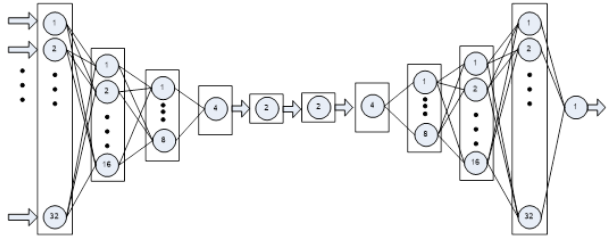
The proposed neural network anomaly detection method is simple and demonstrates fast learning

Data Analysis on the Example of Fraud Detection in Financial Transactions

Indicators of the effectiveness of fraud detecting (Credit-card Fraud) in financial transactions based on data from the Kaggle platform

№	Method name	Performance indicator				
		f1	roc_auc	accuracy	precision	recall
1	KNN	0.705	0.864	0.991	0.643	0.864
2	LOF	0.551	0.576	0.986	0.539	0.576
3	CBLOF	0.653	0.737	0.989	0.614	0.737
4	XGBOD	0.906	0.854	0.998	0.977	0.854
5	AutoEncoder	0.651	0.721	0.989	0.615	0.721
6	*AdaBoostClassifier*	0.903	0.870	0.998	0.941	0.870
7	KNeighborsClassifier	0.891	0.851	0.988	0.922	0.841
NEW result	**Autoencoder Classifier**	**0.925**	**0.870**	**0.998**	**0.999**	**0.870**





Hyperparameters of the neural network Autoencoder Classifier:

- Code size – 2,
- Number of layers – 4,
- Loss function – *MSE,*
- Activation functions – *Relu,*
- Batch_size – 30,
- Optimizer – A*dam*.

Data Analysis Using the Example of Data on Thyroid Gland Disease

UCI
Machine Learning Repository

THYROID DISEASE DATASET - data set on thyroid diseases

Information about the data set
The source thyroid disease dataset (**ann-thyroid**) from the UCI machine learning repository is a classification dataset. It has 3772 training and 3428 test cases. The task is to determine whether the patient referred to the clinic has hypothyroidism. To detect outliers, 3772 training instances with 6 real attributes in total are used – 93 outliers (2.5%).

Indicators of the effectiveness of detecting anomalous data by various methods for the data structure type «THYROID DISEASE DATASET»

Method name	Performance indicator				
	f1	roc_auc	accuracy	precision	recall
Autoencoder Classifier	**0.942**	**0.931**	**0.992**	**0.954**	**0.931**
LocalOutlierFactor	0.701	0.694	0.963	0.710	0.694
AdaBoostClassifier	0.922	0.916	0.990	0.927	0.916
CBLOF	0.770	0.788	0.969	0.755	0.788
XGBOD	0.939	0.928	0.987	0.942	0.928
AutoEncoder	0.716	0.707	0.965	0.725	0.707
KNN	0.737	0.747	0.965	0.729	0.747

II. A Method of Creating «Heat Maps» for Data Analysis

The method of creating «Heat Maps»

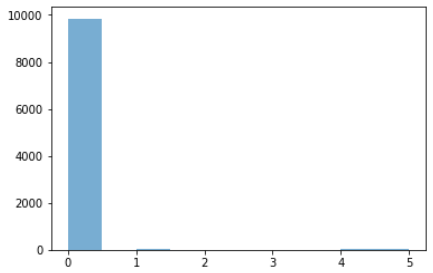
Selection of data processing methods → **Creating a histogram of the results of "voting" methods** → **Histogram analysis and data cleaning threshold selection** → **Separation of normal and abnormal data**

Data processing methods

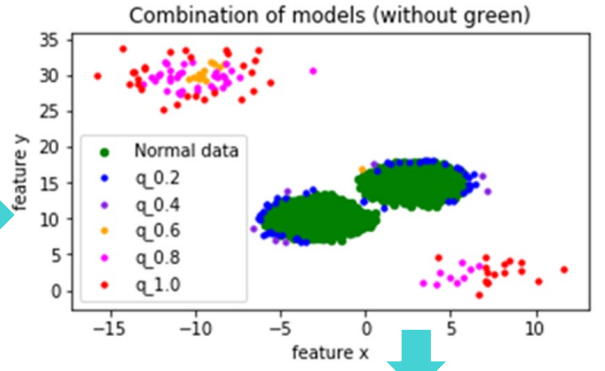
	std	iqr	dbscan	svm	isolation	lof	elliptic	knn	cblo	histogram	autoencoder
0	1	0	1	1	1	1	1	1	1	0	1
1	0	0	1	0	1	1	1	1	1	1	1
2	1	0	1	1	1	1	1	1	1	0	1
3	1	0	1	1	1	1	1	1	1	1	1
4	0	0	1	1	1	1	1	1	1	1	1

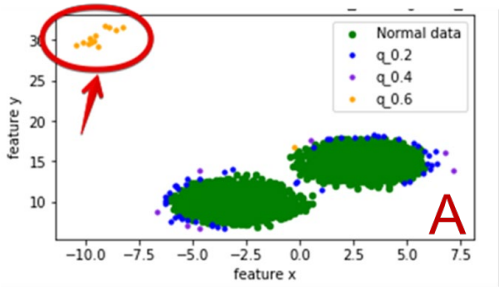
Results of data analysis by various methods

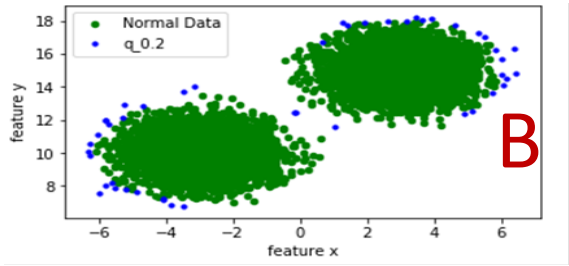
Кількість методів, які назвали дані аномальними	Amount of data
0	9845
1	47
5	44
4	44
3	13
2	7

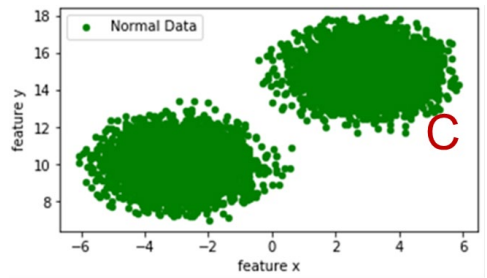


Histogram of the distribution of votes by different methods





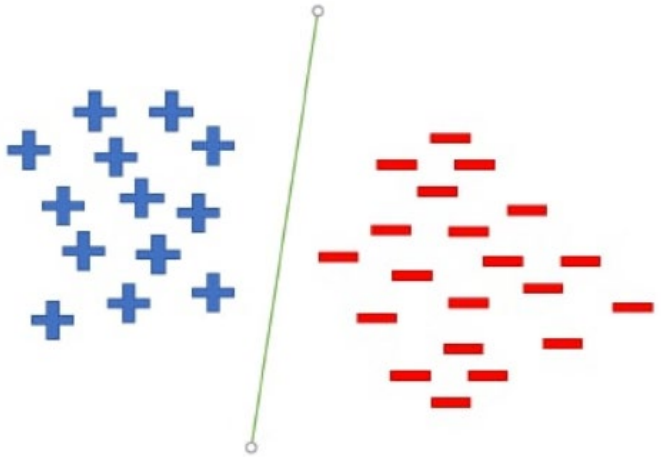




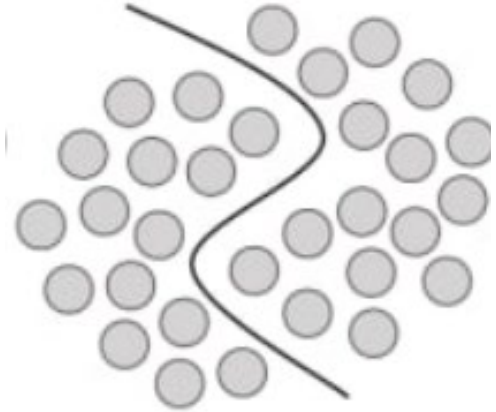
Data sampling without obvious outliers (**A**); with a high probability of normal data (**B**); without abnormal data (**C**)

III. Semi-Automatic Anomaly Detection Method

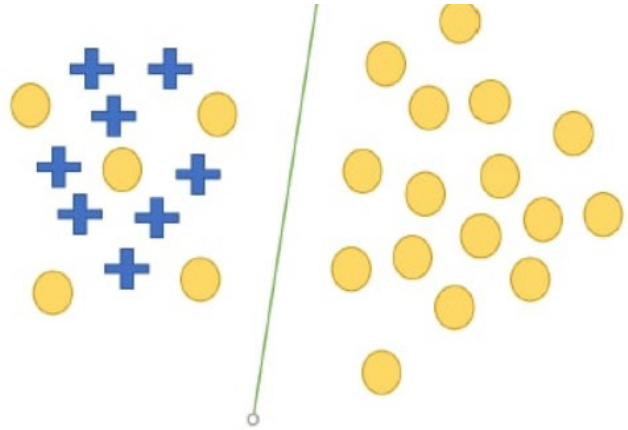
Supervised



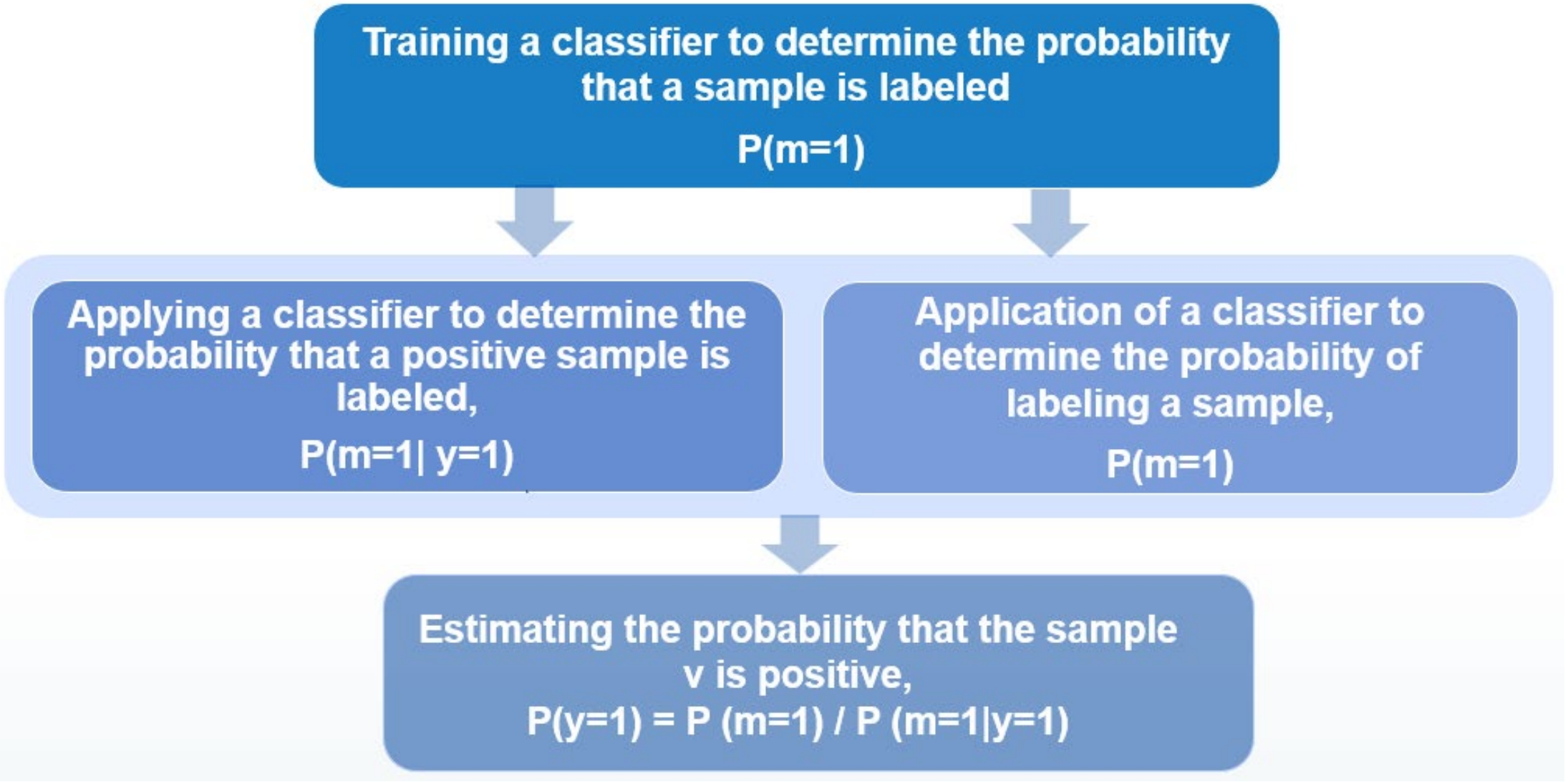
Unsupervised



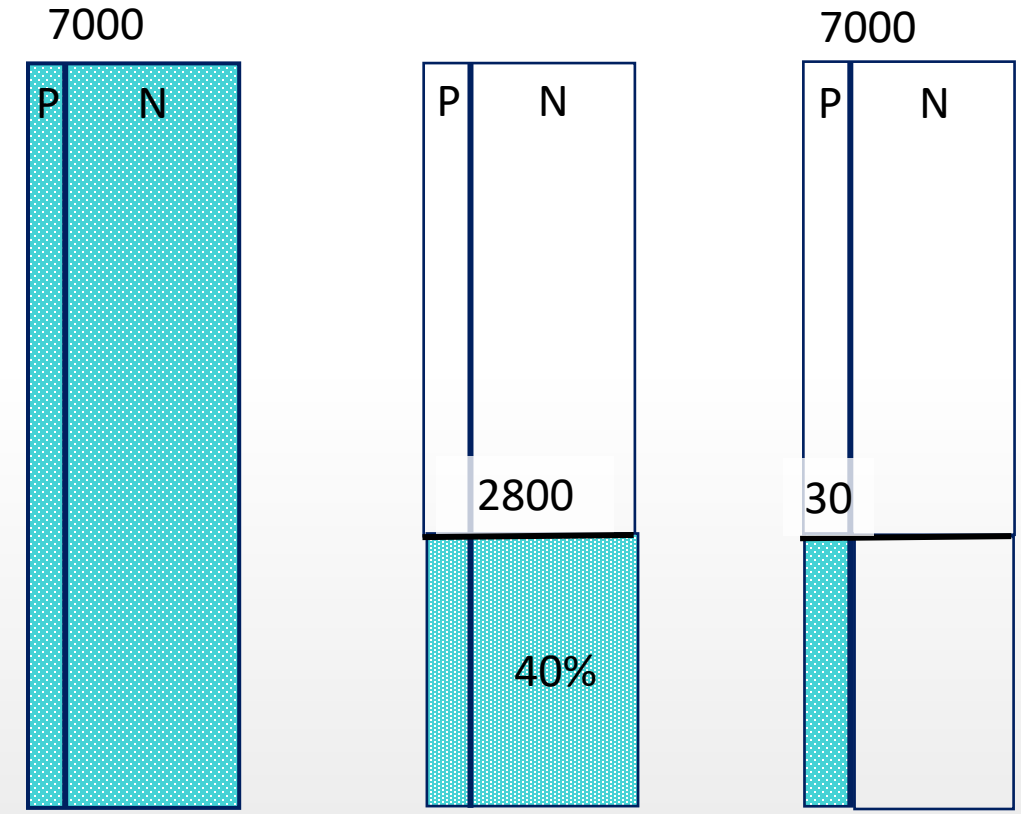
Positive and Unlabeled Data



III. Semi-Automatic Anomaly Detection Method



III. Semi-Automatic Anomaly Detection Method



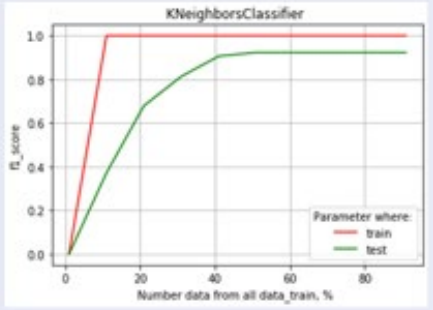
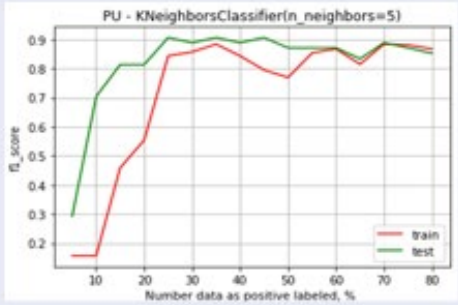
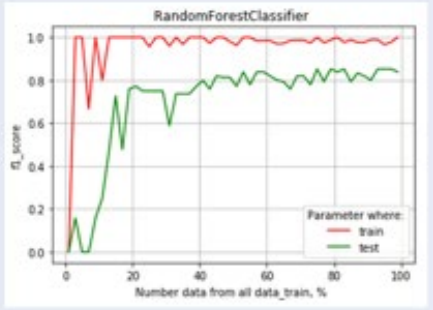
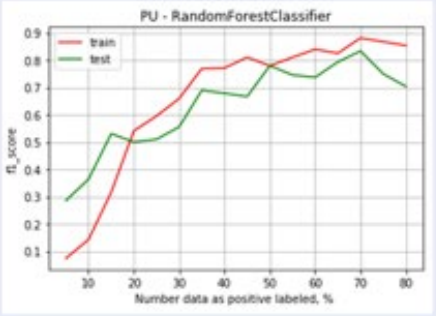
Supervised | Semi-Supervised | Positive and Unlabeled Data

The conducted experiments show that:

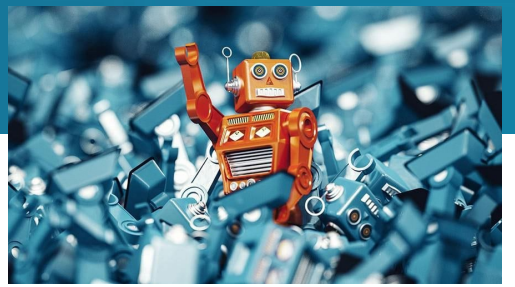
- the traditional Supervised approach required **7000** markings;
- with partial (Semi-Supervised) – approximately 40% of all data, which is **2800** markings;
- and when applying the new labeling method (**Positive and Unlabeled Data**) only data emissions – about **30** data.

III. Semi-Automatic Anomaly Detection Method

The Effectiveness of Anomaly Detection

Partial data labeling	Partial labeling of positive data
KNeighborsClassifier	
	
RandomForestClassifier	
	

Results and Conclusions



- A new **author's method** of improving the quality of abnormal data detection, which is based on the neural network structure, has been developed.
- A method of **iterative data cleaning and creation of "heat maps"** of data based on their probability distribution has been developed.
- A new **semi-automatic learning method** is proposed, which allows using not the entire set of labeled data for solving the *Anomaly Detection* problem, but only the part that satisfies the given accuracy for anomaly detection, which significantly reduces both time and resources in data preparation.