Historically, since 2007, most of the authors of scientific papers on solving the problem of automatic classification of geophysical facies based on logging measurement data used a data set for their studies that is publicly available on the website of the University of Kansas (http://www.people.ku.edu/~gbohling/EECS833/). It is this data set that has been the basis for numerous experiments, collection of quality estimates of the obtained classification results, and comparison of results achieved in various papers. Therefore, in this paper, the data set mentioned above will be described along with its key characteristics and features.

The research area is the Hugoton and Panoma gas fields in southwest Kansas and northwest Oklahoma. These fields make up the largest area in North America where gas is produced. Gas from the Panoma field is produced from the upper seven marine (M)-continental sedimentary cycles of the Permian, Wolfcampian, Council Grove Group , and classification of these rocks is necessary.

The data contains information that the studies were conducted in 10 wells (SHRIMPLIN, ALEXANDER D, SHANKLE, LUKE G U, KIMZEY A, CROSS H CATTLE, NOLAN, RECRUIT F9, NEWBY and CHURCHMAN BIBLE) of the field indicated above. A total of 4,149 measurements of facies properties from these wells were taken at the interval of 0.5 feet (i.e. there are 4149 feature vectors). Thus, the data set contains logging measurements from 10 wells that were labeled with a facies type based on the core data. Sedimentary rocks in this study are divided into 9 facies, classes, which are presented in Table I. However, it is worth noting that the RECRUIT F9 pseudo-well was specifically introduced by the authors of the described data set to improve the classification accuracy of the ninth facies (BS). This was necessary because other wells contain only a small number of BS facies, which makes it difficult to determine them. According to authors of data, a specified number of different facies is the minimum for an accurate representation of the physical variability of the reservoir and the maximum for class division of rocks by their basic petrophysical properties. The facies were distinguished on the basis of the rock type and texture. The rocks of the Panoma field consist of different amounts of the four mineral components – calcite, dolomite, quartz, and clay; the relative proportion of these minerals determines the basic type of rock. In addition, about 80% of the facies are of marine origin, while the remaining 20%, in turn, are of non-marine origin.

TABLE I. DICTIONARY OF FACIES

Facies label	Facies type	Adjacent facies		
SS	Nonmarine sandstone	CSiS		
CSiS	Nonmarine coarse siltstone	SS		FSiS
FSiS	Nonmarine fine siltstone	CSiS		
SiSh	Marine siltstone and shale	MS		
MS	Mudstone	SiSh		WS
WS	Wackestone	MS	D	PS
D	Dolomite	WS		PS
PS	Packstone-grainstone	WS	D	BS
BS	Phylloid-algal bafflestone	D		PS

A characteristic of this data set is that some facies are adjacent since they are close to each other in terms of the similarity of their petrophysical properties. This is the main reason for the gradual mixing of such facies, and therefore, it is rather difficult to attribute them to the appropriate class at the classification (it seems highly probable to make a mistake). Thus, it can be considered that the presence of a classification of facies that is close to the actual

one (classification taking into account adjacent facies or the adjacent facies classification) is satisfactory. In the third column of Table I adjacent facies are indicated for each facies.

Rocks have numerous physical and chemical properties that can be used for classification. The most accessible properties for rocks found in oil and gas wells are those measured by special petrophysical well logging tools dipped into the well after it has been drilled. Table II gives a brief description of the measured properties of facies available in the considered data set. They are used as features for classification by machine learning methods. It should be noted that there are 5 main features in the described data that correspond to the physical properties of rocks (GR, ILD_log10, DeltaPHI, PHIND and PE). These characteristics have been usually recorded for drillable wells since the 1970s, and they are often used by experts for facies classification. In turn, 2 important additional geological features (NM_M, RELPOS) were included in the vector of features characterizing the facies. The prospective significance of these two features for the classification task is due to the fact that some facies are very accurately determined by the values of given measurements.

It is important to mention that there are gaps in these data. There are no measurements of photoelectric effect values for three out of ten wells (ALEXANDER D, KIMZEY A, and RECRUIT F9) since it is not possible to make these measurements universally using well logging tools. Since the value of the photoelectric effect is an important feature for determining facies, it becomes necessary to choose an efficient method for the correct recovery of the missed measurements.

TABLE II. DESCRIPTION OF THE FEATURES

Feature	Feature description	Unit of measure
GR (Gamma ray)	Gamma radiation: measurement of the natural radioactivity of rocks in a well.	API (unit of counting rate at gamma logging of the American Petroleum Institute)
ILD_log10 (Resistivity)	Resistivity logging: measurement of the electrical resistivity of rocks (it is the ability to prevent flow of current).	Ohmmeter (Ωm)
DeltaPHI (Neutron - density porosity difference)	Neutron gamma logging: measurement that correlates with the density of facies.	%
PHIND (Average neutron-density porosity)	Neutron gamma logging: measurement that correlates with the density of facies.	%
PE (Photoelectric effect)	Gamma-gamma logging: photoelectric absorption	eV (electron volt)
Depth	The depth at which the measurements were taken.	Foot (1 ft.= 0.3048 m.)
NM_M (Non-marine / marine indicator)	Indicator of what class facies belongs - marine or non-marine.	"1" is non-marine, "2" is marine —

RELPOS (Relative position)	Index corresponding to the depth at which the measurements were taken (This index starts at 1 and decreases with increasing depth).	—

Significantly, an important characteristic for any data of geophysical studies is the absolute difference of various physical indicators (Fig.1). For example, in the given well log data set, gamma radiation is measured in API; resistivity – in ohm-meters; the photoelectric effect is measured in electron volts. In the light of the above, the comparison of geophysical quantities is incorrect without normalization – it eliminates the difference in the values of these quantities.
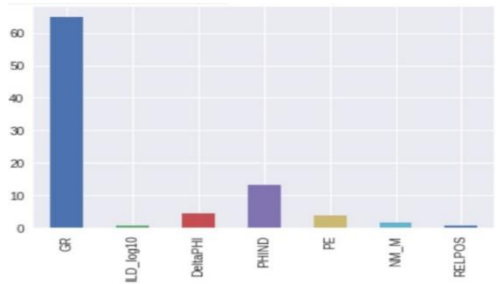

Fig. 1. Mean values of facies features measurements

The complex of measurements obtained during geophysical surveys of a well is usually depicted as a logging diagram (well-logging record). The logging diagram is a set of curves showing changes of the physical properties of the rocks along the wellbore. Fig. 2 is an example of such a diagram for one of the wells of the Panoma field. It was built based on measurements of the facies properties (GR, ILD_log10, DeltaPHI, PHIND, PE) available in the dataset under consideration. The first 5 columns show measurements of these 5 properties (horizontal values) along the SHRIMPLIN wellbore (vertical axis - depth). Each set of measurements at a certain level of depth corresponds to one of 9 facies from Table I (the rightmost column "Facies" of Fig.2). So, researchers can use such logging diagrams as a very useful illustrative way to interpret the results of the classification (if the predicted facies column is depicted along with the existing columns).
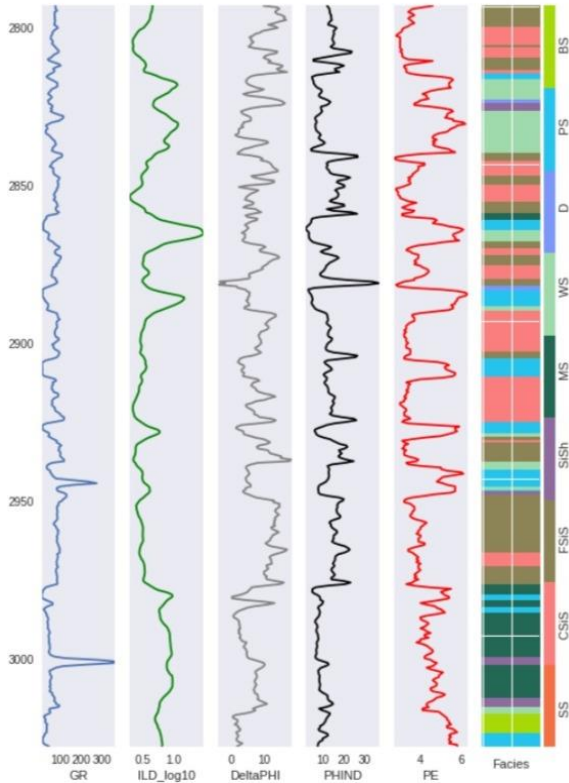

Fig. 2. Logging diagram for well SHRIMPLIN