IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 12, 2020, accepted September 18, 2020, date of publication September 25, 2020, date of current version October 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026315

Bayesian Assessment of Diagnostic Strategy for a Thyroid Nodule Involving a Combination of Clinical Synthetic Features and Molecular Data

ALEKSANDER PŁACZEK[1,2], ALICJA PŁUCIENNIK[1,3], AGNIESZKA KOTECKA-BLICHARZ[4], MICHAŁ JARZĄB[5], AND DARIUSZ MROZEK[2], (Senior Member, IEEE)

[1]Department of Research and Development, WASKO S. A., 44-100 Gliwice, Poland
[2]Department of Applied Informatics, Silesian University of Technology, 44-100 Gliwice, Poland
[3]Department of System Biology and Engineering, Silesian University of Technology, 44-100 Gliwice, Poland
[4]Department of Nuclear Medicine and Endocrine Oncology, Maria Skłodowska-Curie National Research Institute of Oncology, 44-100 Gliwice, Poland
[5]Breast Cancer Unit, Maria Skłodowska-Curie National Research Institute of Oncology, 44-100 Gliwice, Poland

Corresponding author: Aleksander Płaczek (aplaczek@polsl.pl)

This work was supported in part by the National Center for Research and Development Project called Molecular Diagnostics and Imaging in Individualized Therapy for Breast, Thyroid and Prostate Cancer (MILESTONE) through the Program Prevention and Treatment of Civilization Diseases (STRATEGMED) under Contract STRATEGMED2/267398/4/NCBR/2015, and in part by the Silesian University of Technology, Gliwice, Poland, through the Polish Ministry of Science and Higher Education as part of the Implementation Doctorate Program under Contract 10/DW/2017/01/1.

ABSTRACT The use of machine learning has increased over the years, especially in the world of molecular data. Generally, the inference of relationships between features is determined by statistical models. The phenotype (observable clinical characteristics) can result from the expression of the genotype (genetic code) or environmental factors. Molecular datasets have limited information, while supporting clinical data is ambiguous. There are no well-established approaches for combining clinical information with genomic repositories. The genomic tests that are available only use molecular data and give physicians a result which can be integrated clinically. In this article, we present the strategy where clinical data, regardless of its limitations, is combined in one predictive model with molecular features. We predict the risk of malignancy in the thyroid nodules based on the results of fine-needle aspiration biopsy and expression of selected genes. We utilize a Bayesian network (BN) framework to discover relationships between molecular features and assess the impact of added clinical data quality on the performance of the chosen gene set. Bayesian network offering both prognostic and diagnostic perspectives is a perfect non-parametric technique for feature selection, feature extraction, and prediction purposes. We show that certain clinical factors could work as a synthetic feature and provide predictive abilities beyond what genes alone can offer. The experimental results demonstrate a higher performance of predictive models based on molecular and clinical data than when using only molecular data. We also explain why, one should consider the source of clinical data, but be aware of the quality of variables.

INDEX TERMS Bayesian networks, feature integration, synthetic features, Markov blankets, Quality of features, thyroid cancer, Bethesda classification.

I. INTRODUCTION

In 2018, according to the World Health Organization, the number of new cases of thyroid cancer (TC) in Poland was more than 3,600. Since the beginning of the 21st century, this number has increased every year. The thyroid cancer related mortality rate is low and ranges between 0.4-0.6 deaths/100,000 people [1]. An independent work [2], showed that from a good prognosis, the main problem we have to face in the diagnostic process results from the fact that malignancy in the thyroid nodule only occurs in about 5% of patients who have them. Therefore, proper stratification of the risk of malignancy in the thyroid nodule remains the major challenge. Precise stratification would allow the

The associate editor coordinating the review of this manuscript and approving it for publication was Sabu M. Thampi.

establishment of a final diagnosis and enable the selection of those patients who really need surgical treatment. Unnecessary surgeries result in lifelong drug administration and a reduction in the quality of life.

Diagnostic difficulties in the management of patients with thyroid nodules and, in some cases, the ambiguity of cytological findings, have contributed to the development of new molecular tools. Continuous progress in technologies that allow simultaneous testing of multiple genes results in lower molecular testing costs and an increased number of studies on molecular markers. A recently published paper related to TC raises the role of molecular classifiers [3]. At the same time, the supportive role of clinical data is underestimated. Classifiers that are available for thyroid nodules operate only on molecular data without including clinical features in the stratification risk algorithm. This means that these factors need a separate medical evaluation by a physician. According to National Thyroid Associations guidelines, results must always be integrated with the patient's clinical status, appearance of the ultrasound scan on the thyroid nodule and its cytological analysis [4], [5]. According to American Thyroid Association guidelines, molecular tests may be suggested as additional risk stratification tools in the categories mentioned. In the Guidelines of the Polish National Societies Diagnostics and Treatment of Thyroid Carcinoma (GPDT) 2018 Update, the molecular test is suggested as an additional tool that helps to distinguish between benign and malignant nodules. Such an examination is recommended only in the reference medical centers experienced in the molecular investigation. At the same time, research aiming to develop cost-effective molecular testing, affordable in Poland, are mentioned.

Combining features from different sources is a widely used technique [6]–[9], e.g. using meta-analysis as an alternative approach to parameter learning from data or expert knowledge [10]. Bayesian networks have also been frequently used as predictive models in a different type of cancer [11]–[13]. Fenton and Neil [12] presented a Bayesian network as a method preferred by users who do not believe in ''black box'' algorithms. A graphical representation of the probability distribution over a large number of casual variables allows the inference to be followed from evidence to the event. Another desirable property is the ability to set one's own prior probability to a particular independent variable just before calculation: a medic's risk assessment. Zhao and Weng [7] demonstrated how combining Electronic Health Records (EHRs) and seemingly independent sources like PubMed databases can impact the performance of the BNs evaluated. Additional information allowed them to calculate the weight, i.e. importance, of the nodes which increased the Area Under a Curve metric (AUC) by about ten percent points above the conventional BN, outshining other methods. This combination is essential in the area of medicine due to inconclusive information being obtained within the diagnostic process when data from a single source is analyzed.

None of the recently developed molecular tests which support physicians in the management of thyroid nodules with indeterminate cytopathology, use clinical data in the malignancy risk stratification algorithm [14]–[17]. Finally, there are not many studies available which have compared the models using the same variables but with different qualities.

This article aims to show that some clinical features can form synthetic features and, despite quality issues, they may support genes in the prediction of the risk of malignancy in thyroid nodule. We assume that they can provide general information, which, in some cases, may not be consistent with molecular evidence. The impact of such a lack of matching has been described and discussed. We compared both the strategies mentioned above to assess which is the better solution for medics. This study is part of a more comprehensive project called MILESTONE, in which different feature sized classifiers are considered, and an intuitive tool to compare their results is required. To build such a tool, we utilized Bayesian networks, which offer both prognostic and diagnostic views on the node's relationships while keeping the same joint probability distribution.

A. CLINICAL ANALYSIS BACKGROUND

The initial management of a patient with a thyroid nodule must be focused on clinical risk factor analysis. According to the National Thyroid Association's guidelines, clinical risk factors such as a thyroid nodule bigger than 4 cm, a history of thyroid problems in the family and a history of neck radiation should be considered as indicators of a higher pre-surgical risk of malignancy. According to GPDT, thyroid nodule diagnosis in patients younger than 20 and older than 60 may also increase the risk of malignancy. Furthermore, lymph nodes that have been confirmed to be cancerous, occurrence of distant metastasis, rapid thyroid nodule enlargement, and symptoms such as swallowing difficulties or voice change also significantly increase the probability of malignancy in the thyroid nodule [5], [18].

In the diagnostic process of the thyroid nodule, an ultrasonography examination is the first step of imaging, after clinical assessment. There are several Thyroid Imaging and Reporting Data Systems (TIRADS) applied in clinical practice. Depending on the estimated risk for a particular image, the threshold for the size of the thyroid nodule is set for further evaluation. According to the ultrasound reporting system proposed by the American College of Radiology (ACR-TIRADS), a biopsy of nodules deemed highly suspicious is recommended if they are 1 cm or greater in size. ACR-TIRADS advocate the biopsy of a nodule with a low risk of malignancy when the nodule measures 2.5 cm or more [19].

Fine needle aspiration biopsy (FNAB) of the thyroid nodule produces material for cytological analysis. Cytopathology provides clinically useful data in the form of the widely used Bethesda System for Reporting Thyroid Cytopathology (TBSRTC, Bethesda), comprising of 6 categories assigned with different malignancy risks, which drives

TABLE 1. Empirical risk rates in Poland based only on data gathered and published by NCI. Source: The 2017 bethesda system for reporting thyroid cytopathology, Table 2 [18]. *lack of Polish data - data given in the table are NCI data.

Category	Recommended terminology	The Risk of malignancy	The Risk of malignancy considering NIFTP as postoperative outcome	The risk of malignancy in Polish patients	Cytological diagnoses included in a particular category and other comments
I	Nondiagnostic or unsatisfactory	5-10	5-10	5-10%*	Clinical context should be considered
II	Benign	0-3	0-3	<1%*	Nodular goitre Thyroiditis, including chronic inflammations Hyperplastic nodule Colloid nodule (lots of colloid, sufficient cellularity) Cytological findings suggest colloid nodule (lots of colloid, insufficient cellularity) Thyroid cyst
III	Atypia of undetermined significance (AUS) or Follicular lesion of undetermined significance	$\sim 10-30$	6-18	2.4-5.2%	This category should be used in rare cases when it is not possible to state a precise cytological diagnosis
IV	Follicular neoplasm or Suspicious for a follicular neoplasm	25-40	10-40	8.2-19%	At least 25% of lesions belonging to this category are not neoplastic tumors (hyperplastic nodules, inflammation). This category should not be diagnosed when nuclear features of papillare thyroid cancer are present
V	Suspicious for malignancy	50-75	45-60	75%	This category involves: — papillary thyroid cancer — medullary thyroid cancer — lymphoma — metastatic carcinoma — anaplastic thyroid cancer/vascular sarcoma due to the presence of necrotic tissues
VI	Malignant	97-99	94-96	95-100%*	This category involves: — papillary thyroid cancer — medullary thyroid cancer — lymphoma — metastatic carcinoma — anaplastic thyroid cancer/ vascular sarcoma

patient management [20], [21]. Unfortunately, about 20% of cytopathology results are inconclusive falling in one of the indeterminate categories, i.e. Bethesda III (B3) or IV (B4), with the risk of malignancy varying widely among populations [5], [18], [21]. Rates for part of the Polish population were published in GPDT, as shown in Table 1.

Decisions about surgery should be taken individually based on risk assessment and the experience of the center. In the case where a unit does not have much experience, surgery will be promoted for safety reasons. Results that have been reported on the use of gene-expression classifiers in the diagnostics of indeterminate thyroid nodules are very promising [2]. However, in countries where molecular tests are available, the decisions concerning surgical treatment is based on a cytological assessment with molecular analysis in a supportive role.

Clinical features are also used in risk stratification for TC and may affect the patient's prognosis. One such feature is the patient's age. Survival rate is reduced for middle-aged patients [22], and gets progressively worse. The role of age as a factor has also been confirmed by Zaydfudim *et al.* [23], who showed that the age of 45 was identified as a cutoff point after which lymph node disease impacts survival rate. Recent validation from many international institutions [24] resulted in a change of risk stratification, with the age cutoff for thyroid cancer changing from 45 to 55 in the American Joint Committee on Cancer's (AJCC) staging system. The initial conclusion from these studies is that age should be taken into account during the analysis as to whether surgery is necessary; the research by Kelly *et al.* [25], confirmed this. However, this does not mean that this is the direct cause of a lesion's malignancy (i.e., risk stratification in the thyroid nodule), but it is an indicator that the prognosis of malignancy is worse when the nodule is malignant. These risk factors are related to the population and the Human Development Index [26], [27].

Recommendations about paying attention to clinical features are also included in the Polish GPDT. The same papers report that more women are affected by TC, but the stage of malignancy is higher for men. This was the reason why we took into account those particular features as potential factors in the risk stratification of the thyroid nodule.

B. MOLECULAR ANALYSIS BACKGROUND

Over the last decade, some molecular tests have been developed. These are based on the analysis of differentially expressed genes. One of these tests is the test presented in the paper [15], in which local cytopathological reports were used (primary Bethesda) and, if necessary, reclassified by three (3) experienced specialists (revised Bethesda). The authors of this article also stated that in 14% of revised cases, there was a disagreement in the final assessment. However, no further information about the relationship between local recognition and revised assessment was reported, i.e. changes in scoring. This classifier was finally developed based solely on gene expression data within the subset of patients who were selected by cytopathology. Other tests [14], [16] that just focused on indeterminate lesions did not integrate clinical features into the algorithm.

A comparison of the main characteristics of the tests [4] and their analysis by independent evaluation [28] show a significant difference between the performance reported by the test creators and what was actually achieved in the evaluation. It stands to reason that in real-life it is harder to select a perfect group of patients, and the test would be carried out in a less meticulous way than in a clinical trial. So, when there is an uncertainty about the classifying criteria, the classification performance may be affected more. Independent studies highlighted that the results of classifiers could also vary depending on the center's experience. Both aspects are related to the quality of the cytological assessment and may impact the diagnostic process. When Bethesda categorization is carried out by a pathologist inexperienced in the area of TC (especially in B3 or B4 samples), clinicians may have potential problems with integrations and rejections.

In this study, we took a gene set from a published gene expression classifier, as described by Chudova [29], which was proved to distinguish between benign and malignant lesions in indeterminate nodules.

C. THE IDEA BEHIND THIS ARTICLE

Realistically, indeterminate lesions are probably the target population most suitable for molecular classifiers. However, this group is not clearly defined and is not heterogeneous in its nature. The consulting pathologist provides a vast amount of data which could potentially be taken into account when providing the feedback based on molecular test results. Thus, the idea behind our analysis was to integrate the probability based on prior knowledge resulting from clinical classification into decision-making when the molecular test result is available. This process is carried out by every physician confronted with having to make a decision about his patient; we tried to incorporate this formally into the process.

Trying to find out the impact of adding clinical features to the previously discovered set of distinguishing genes, we considered some additional questions:

1) How can combining clinical features with genes change the basic molecular set's performance?
2) Is there a difference between the impact of locally performed tumor lesion assessment and the impact of expert judgment on the classification performance ?
3) What if the Bethesda assessment is mismatched? Could this result in the wrong interpretation and influence test outcomes?

We were also motivated by the problems with the accessibility of molecular diagnostics in Poland, where the usage of commercial molecular tests is costly. The decision about testing is made based on clinical evidence and the Bethesda category. In practice, features verified accurately during test development and the validation procedure can be mismatched. The Bethesda categorization requires experience in the area of TC and much practice with different samples. However, a suitably qualified person with the relevant experience may not be available; in this case some patients may be rejected from molecular testing because of mistakes made by a less-experienced specialist. Finally, even if the molecular test is done, and the result is opposite to the cytological assessment, the clinician's decision may be difficult.

It was assumed that for any reason, the feature selection process had not taken clinical variables into account, when classifier [29] was developed. Consequently, the study started with molecular features that distinguish between benign and malignant lesions of the thyroid.

The Bayesian network framework was chosen to build the reference model and all of the rest of the combined models [30]. This is a great tool to reduce initial dimensionality and to discover the network of dependencies among a given set of variables. The knowledge discovered about the uncertain domain (dependency between gene expressions and malignancy) is represented in the form of a directed acyclic graph (DAG). Visual representation facilitates the ability to distinguish cause and effect from the correlation. It is a very intuitive framework to analyze relationships between the variables taken into account in the unsupervised analysis.

II. BAYESIAN NETWORK BACKGROUND

Bayesian networks have become very popular in recent years. Data is represented in the form of a DAG that presents conditional dependencies among a set of variables. The graph consists of nodes representing features and arcs representing known or discovered relations between variables. The structure of the graph can be built manually based on expert knowledge or learned from data using constraint-based or score-based learning algorithms.

For the DAGs in this article, the focus was only on three properties:

- d-separation (separation of nodes in a directed graph),
- the Markov blanket,
- the strength of arc.

Only discrete variables were used. Each discrete variable can have a finite number of states. They can be connected directly or indirectly (via another node), but in both cases, a particular state of node X can cause a change in the probability distribution of node Y. Data can only be introduced into some of

these, marginalizing the rest, and the probability of a specific state of the variable of interest calculated. The whole network encodes a global joint probability distribution over the nodes. By using the d-separation property, a joint probability distribution can be replaced by the product of conditional probability distributions for each variable, as presented in Equation 1. This reduces the need for computing resources and limits the number of variables that need to be observed.

$$P_x(X) = \prod_{i=1}^{n} P_{x_i}(X_i | \prod x_i), \tag{1}$$

where

X_i is the variable of interest in the factorization step,

$\prod x_i$ is the set of the parents ofX_i.

To discover whether two nodes (e.g. X and Y) become independent after introducing data into the network, all paths between those variables must be checked. Variables may block communication when they are instantiated and they are:

- in the middle of the path between X and Y (serial connection),
- parents of both (diverging connection),
- children of both (converging connection).

Definition 1 (D-Separation): According to the definition provided in [31], two variables are d-separated if, for every path between them, there is an intermediate variable Z (distinct from X and Y) such that one of the following conditions is true:

- the connection including all variables is serial or diverging, and Z is instantiated, or
- the connection is converging, and neither Z nor any of Z's descendants have received evidence (data)

From definition 1, we can easily deduct that there are nodes in the nearest neighborhood of each variable, which, when instantiated, make this variable practically insensitive to changes in the rest of the network. This particular set of nodes is called the Markov blanket.

Definition 2 (The Markov Blanket): According to the definition in [32] the Markov blanket of node A is the minimal subset of nodes such that:

$$MB = A \perp\!\!\!\perp p \quad V - S - A | S \tag{2}$$

where

p : means probabilistic independence,

V : is the set of all variables,

S : is the subset of V.

The Markov blanket property can also be described in probabilistic form:

$$P(A|MB(A), V \setminus \{MB(A), A\}) = P(A|MB(A)) \tag{3}$$

where

MB : means the set of nodes standing for the Markov blanket,

V : is the set of all variables,

Regarding graphical separation, the Markov blanket consists of parents of node A, their children, and spouses. From Equation 3, one can see that to predict the probability of a particular state of variable A, only states of the Markov blanket nodes must be entered as evidence. The rest of the nodes might be marginalized.

The results of structure learning and parameter learning should be validated before using a BN for inference in medicine. One of the methods used to do this is *bootstrap re-sampling* [33]. It is necessary to do this validation because the structure learning process is sensitive to data, especially in the vicinity of arcs, i.e., existence and direction. The bootstrap re-sampling method takes a subset of samples and learns the structure. Then the model takes average values, and, as a result, it provides two features of BN: strength (the number of times when the arc was present in the learned networks modulo its direction) and direction. In a perfect DAG, the arc strength is the confidence measure that this arc (relation) exists. Dependent variables are connected via a direct arc in the network. Conditionally independent variables are connected through the path (a sequence of nodes and arcs).

Bayesian networks are reported as a great tool in individual-level risk prediction, can be easily extended into decision models by incorporating a decision node [34], and are used for feature extraction, including the area of medicine [35]. Incorporating nodes from different sources into one network allows the researcher to identify the causal structure between variables and assess which features are related to the predicted variable (e.g. malignancy node). On the one hand, the number of molecular features obtained from high throughput molecular experiments (like microarrays or sequencing, protein mass spectrometry, quantity PCR) can be huge. On the other hand, there are only a few clinical features. The strength of the relationship between the clinical predictor variable and predicted node can be omitted, during the process of learning the network structure from combined molecular and clinical data. Thus, we proposed using the Bayesian network to predict the clinical risk of malignancy in a thyroid nodule and then incorporating the result of this prediction as a synthetic feature into the molecular Bayesian network. We presented that such an approach allows this variable to stand for an essential feature and to enrich other methods of malignancy prediction.

III. MATERIALS AND METHODS

Our research process consisted of several steps. A Bayesian reference network was learned just from clinical data stored in hospital databases. Next, networks were built, learned from the molecular and combined sets of features, respectively. Predictions obtained from networks by conditional queries were compared in order to evaluate whether the quality of input data significantly changes diagnostic reasoning. Attempts were made to predict the malignancy probability by using the level of expression of specific genes or Bethesda categories as evidence.

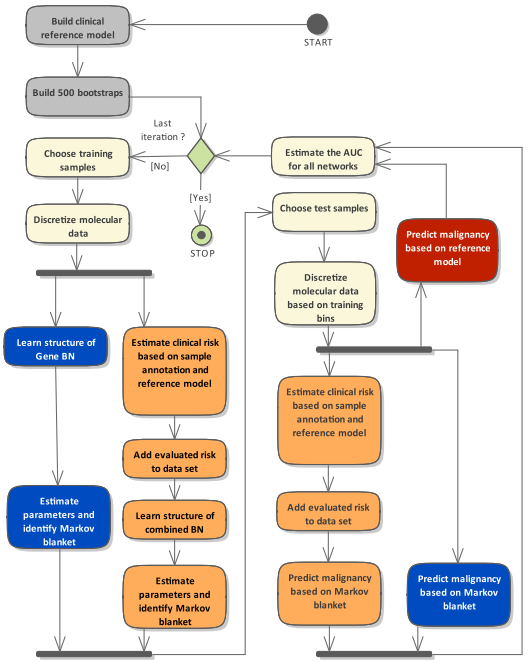


FIGURE 1. Overview of the research process. Colors correspond to the ones used in further analysis, apart from grey, which depicts preparation steps, and sandy, which depicts common actions. Details of each step are shown in the following subsections. The same process is provided for primary and revised Bethesda, prepared as annotations for every sample.

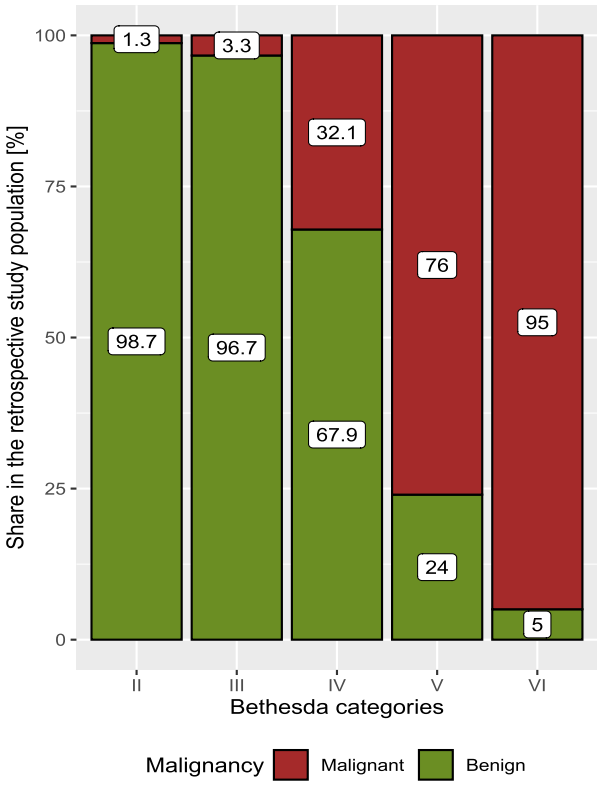


FIGURE 2. Malignancy distribution based on the 2007 - 2017 period among Polish patients treated in the Silesian branch of NCI.

The figures presented in this article do not contain the names of the molecular features due to the pending publication of the set of features for the new TC molecular test.

A. DATASETS

The clinical data samples was collected from the database shared by the National Cancer Institute (NCI). The 8,000 thyroid FNAB reports from the period of 2007 to 2017 were analyzed retrospectively. Depersonalized data was extracted from the database where they remains in an unstructured format. Problems that had to be dealt with were dealing with changing terminology over the years, the lack of the Bethesda categorization in many older reports, and surgery decisions described in native languages within medical documentation. The RetroNGSC classifier (95% k-fold accuracy) for retrospective data [36] was designed, implemented, and used to get fully categorized biopsy reports. Next, medical documentation was examined to find out who had had surgery and what their histopathological results were. This enabled the labelling of malignant or benign to be made to virtually all biopsies. Two crucial assumptions were made: firstly, patients with a Bethesda category of B2 and who were not due to visit for the next two years, had their class set to benign, despite the lack of surgery and existence of a histopathological report. Secondly, for patients with Bethesda category B5 or B6, and without information about their histopathological results from a biopsy, the class was set to malignant; either the patient had died or had not agreed to surgery. These assumptions resulted from the fact that not every surgery was performed in NCI. The results achieved were presented in Figure 2 and are consistent with the values mentioned for the Polish data in the guideline [18], with exception for the B3 category.

Molecular data was collected by the National Cancer Institute (NCI) for the MILESTONE project during the FNAB of thyroid nodules. Gene expression was measured using Affymetrix HTA2.0 arrays. The analyzed data set comprises of 198 arrays (75 malignant and 123 benign) pre-processed by the Department of System Biology and Engineering, of the Silesian University of Technology. Data was pre-prepared with the Aroma tool in the R/Bioconductor environment [37] with the RMA background correction and quantile normalization method. Probes were mapped on ENTREZ genes using the custom Chip Definition File provided by [38]. From normalized and annotated data, features extracted were presented as molecular markers for thyroid malignancy in [29]. The transcripts are involved in a variety of cellular and biological processes. Out of the 167 transcripts which are listed as part of the gene expression classifier, 163 were available in our data set.

B. BUILDING THE CLINICAL REFERENCE MODEL

The clinical network was learned directly from the NCI database. The distribution captured by the network was compared with the existing literature to ensure that predictions based on this model are reliable. Figure 2 depicts a histogram of the Bethesda score either written directly by a specialist or

deduced by our algorithm mentioned below, indirectly based on the cytological report's content.

In 2017 the Bethesda system was revised after the reclassification of thyroid neoplasm previously considered as malignant to benign [39]. In this category the risk has therefore become lower. In the results for this article, the occurrence of malignant cases were more similar to the global risk before this change, since the results were based on records assessed and stored in the database before this change was made. From the dataset, records were excluded where, despite existing medical documentation, a malignancy state could not be established by using text mining techniques alone.

The following features were considered:

- age,
- sex,
- size of lesion,
- shape,
- Bethesda category,
- the existence of multiple lesions,
- echogenicity,
- additional information about affected lymph nodes.

Amongst all these variables, age and size were continuous and, before learning networks, these were discretized, as presented in Table 2. Considering the results of research about the impact of clinical features on a risk stratification in TC, we arbitrarily utilized information about recent changes in age stratification cutoff and ACR TI-RADS recommendations to build intervals.

TABLE 2. Feature bins after discretization.

Bin	Patient's Age	Size of lesion
1	25- (<= 25 years)	XS (<=5 mm)
2	45- (>25, <=45)	S (>5 mm, <=10 mm)
3	55- (>45, <=55)	M (>10 mm, <=15 mm)
4	75- (>55, <=75)	L (>15 mm, <=25 mm)
5	75+ (older than 75 years)	XL (>25 mm, <=40 mm)
6		XXL (bigger than 4 cm)

Figure 3 presents the Bayesian diagnostic network learned from this data and used in further analysis as a clinical reference model.

All selected variables might be observed during laboratory or physical examinations. *Size, lymph node metastases (Lymphs), and echogenicity (Echo)* are determined from thyroid ultrasound, and *Bethesda (Beth)* is the state of tissue interpretation under a microscope. These features could not be treated as a separate risk, which can cause malignancy, but their specific values could suggest an increased probability of malignancy. Since not all variables were collected for samples used in our research, we decided to use only: *age, sex, Bethesda (Beth)* and *size*. Providing them as pieces of evidence and marginalizing other variables, we estimated the clinical risk of malignancy for every sample (reported as *Clinical Risk*).

Figure 4 presents the characteristics of the samples' clinical features.

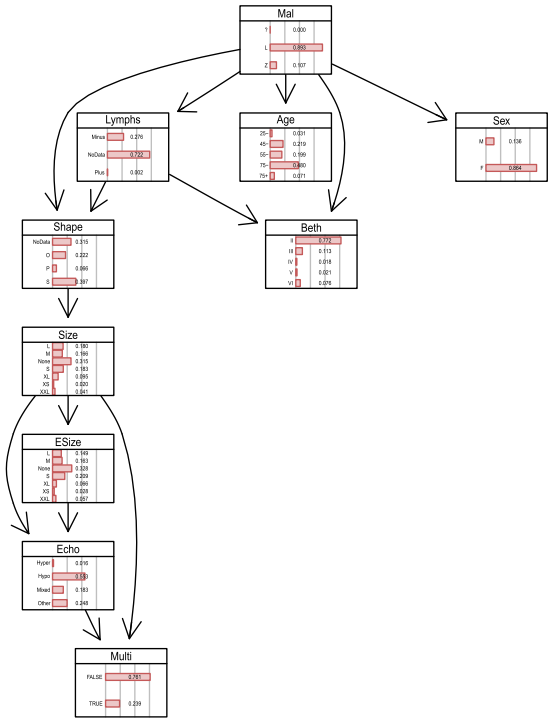


FIGURE 3. The clinical Bayesian network's diagnostic perspective presents relations between clinical features and malignancy events learned from the database over a ten-year period. The probability of malignancy is called *Clinical Risk*.

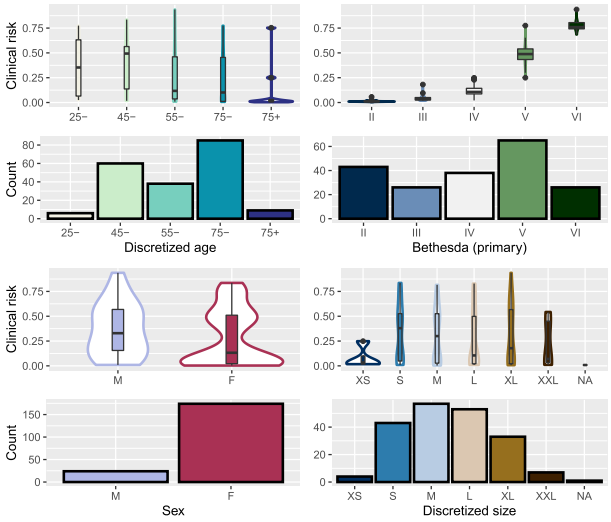


FIGURE 4. The characteristics of samples from the perspective of clinical features. All variables are discrete. There is a histogram and a box plot for the clinical risk presented for each feature.

The risk estimated from BN was a continuous variable; therefore it had to be discretized before further calculations. For this purpose we used the affinity propagation (AP) clustering algorithm [40], [41]. The AP was applied to find a set of similar distributions using gene-expression data of patients as samples. This is a high-speed and straightforward algorithm. It takes a collection of real-valued similarities between data

points as an input and transmits real-valued messages along the edges of the network until a good set of exemplars comes up. We built a similarity matrix between patient samples using a Jaccard distance metric [42]. Each patient (sample) was described by a four element vector (four probabilities): the risk of malignancy (RofM) given the patient's sex, age, nodule size, and Bethesda score. All values were estimated using the Bayesian clinical reference model. This standardized the parameters and allowed them to discover exemplars (members of the input set that are representative of clusters). The most desirable property of this method is the lack of initial setting of the number of expected clusters. We used exemplars to create bins for the discrete *Clinical Risk* variable without any prior assumptions.

C. MOLECULAR DATA DISCRETIZATION

Gene expression data was continuous. Due to the non-normal distribution visible in the data, we had to convert continuous variables into discrete ones. Many discretization techniques are dedicated to gene-expression data [43], [44] trying to achieve a trade-off between the loss of information and cost of computation. An unsupervised discretization approach was used to discover unknown relationships between features.

Firstly, we standardized gene-expression data by subtracting the mean from each value and dividing by the standard deviation. Our goal was to assign the same discrete values to all similar variables to ensure that only the different levels of expression of various genes were distinguished between benign and malignant tumors. Then, we divided each gene expression level into three intervals using the k-means algorithm. We created three bins: less, normal and more.

We also checked the binary level of discretization. However, a comparison of the partial results showed that three discrete values should allow us to achieve better results.

D. MOLECULAR SAMPLE ANNOTATIONS

Each molecular sample was annotated. Sex, age, and size of bi-opted nodules were recorded. For the Bethesda variable, two values were collected. The first one, which we called *Primary Bethesda*, was set by a local specialist during the local cytological assessment prior to sending tissues (or microscope slides) to the reference center. The second one, which we called *Revised Bethesda*, was the result of reclassification or confirmation based on aspirated material or slides taken by at least two pathologists in the reference center.

The approach used assumes that NCI specialists are more experienced in the area of TC than local medics, and their assessments are of better quality. Their assessments should be more consistent with molecular findings. Both values were used independently in our research to compare their impact when added to the molecular dataset. When estimating the clinical risk of malignancy from the clinical reference model, these values were used as data together with the other three features (sex, age, and size of biopted nodule) to calculate the probability of event MAL = 'M' (the state of Malignancy node representing a malignant tumor). The results (the estimated *Clinical Risk*) were integrated with the rest of the molecular variables and the BNs were allowed to select the most appropriate features from the data set.

E. BOOTSTRAP APPROACH

Only a small number of molecular samples were available (198), and the initial analysis showed that the data was not entirely consistent with the population distribution; comparing a clinical reference histogram with one for all annotations collected for the molecular dataset. The dataset was imbalanced. The imbalanced ratio was 1.64; more benign samples than malignant. The bootstrap technique [33] was used to evaluate the performance of each Bayesian network classifier.

The same clinical reference model was used in all bootstrap tests. The clinical risk was evaluated and the result then compared to a class variable (malignant or not) based on the annotations of the test samples. That was the baseline used, i.e. only a risk assessment on clinical data, and at the same time, the value was added to each sample in the combined networks.

500 iterations were planned, sampling 200 records each time with replacements. Those of the records that were not sampled to particular iteration of bootstrap, constituted the testing dataset of this iteration.

During each iteration, the BN structure was trained, and for this the parameters were estimated using bootstrapped instances corresponding to three data sets: molecular data, the same data combined with the clinical risk based on *Primary Bethesda* and the last set combined with the clinical risk based on *Revised Bethesda*. Every network was learned using a TABU search with a multi-nominal log-likelihood score. The TABU search is a search technique based on the greedy search algorithm. It tries to avoid getting stuck in local minima by selecting a network with a slightly worse score than the optimal. The score is equivalent to the entropy measure, which is, in general, the measure of uncertainty for the researcher's assumptions and enables making precise statements because it is a measure of how well the network can predict the class state [31]. To avoid building an extensive network, we limited the number of each parent node to three. The most crucial node in each network was the one representing malignancy.

F. FEATURE SELECTION - THE MARKOV BLANKETS

In this article, the focus was on the Markov blanket (MB) of the Malignancy Node (MAL) in order to reduce the complexity of queries as presented in Equation 3. The property used was that when all nodes in the MB of MAL are instantiated, the changes in the certainty of nodes in the rest of the network minimally influence MAL. Complete data samples were collected in preparation for any set of nodes that might be discovered. For each BN, by averaging the Malignancy Node's Markov blanket from bootstrap tests, features were extracted that existed in MBs more than 50 times, no matter if they acted as a parent, child or spouse node. Figure 5 shows the highest-ranked features for each classifier.

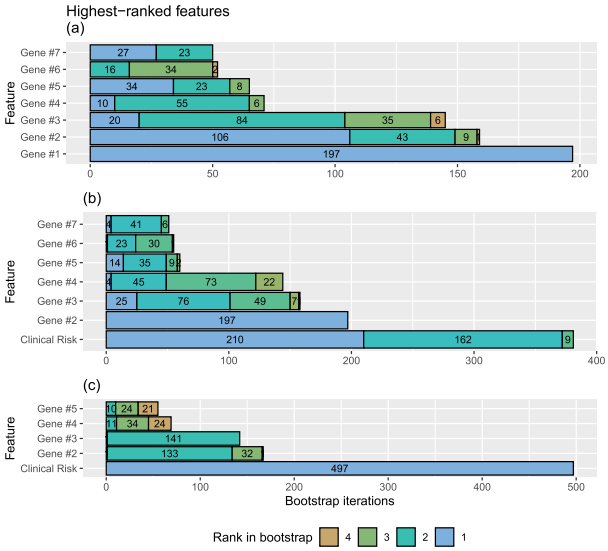


FIGURE 5. **Highest-ranked features obtained from a bootstrap run for various variants of the malignancy risk classifiers relying on (a) only molecular data – GBN, (b) molecular data combined with primary Bethesda – PMBN, (c) molecular data combined with revised Bethesda – RMBN. Colors correspond to a position in the ranking for every iteration. Gene identifiers in the plot (b) and (c) are the same genes as in the plot (a).**

The feature's position was calculated based on the strength of arc drawn between the node representing Malignancy (MAL) and the node representing analyzed feature (e.g. Gene#1) belonging to the MB of MAL. Values of strength were normalized using the min-max formula, then a weight of 1 was added to all direct connections to MAL. Edges between spouse nodes and children of MAL were given a weight of 0.5. After summing up the data grouped by the nodes, the rankings of nodes were determined in descending order.

G. PERFORMANCE METRICS

Before summarizing the results obtained, it is important to emphasize that the dataset used was imbalanced and that most of the samples were benign. This might result in quite good accuracy for each model, especially since the gene set used was initially discovered during the development of the rule-out classifier (which prefers negative samples to positive ones).

There is some discussion in the literature about which metrics to use when comparing different learning algorithms. Accuracy only takes into account correct classifications, and if the negative class is wider, it has a significant impact on the result; it is sensitive to class imbalance. Since this is the case here, accuracy was not the best metric. Many researchers prefer AUC in order to compare models. The AUC is a great metric enabling the overall classifier's ability to distinguish between two classes compared to random guesses. However, there are some limitations when comparing different learning algorithms [45]. Nevertheless, in this case, all the networks are learned from the data, and no prior distributions are set.

BNs are used indirectly as a feature selection tool limiting features of interest to MB of MAL. In this way, a different set of nodes and parameters in each model were obtained, and all were discovered during the training process. The AUC metric was used to compare models in each bootstrap iteration and it was decided that the Matthews correlation coefficient (MCC) should be used instead of accuracy. Full symmetry is the crucial property of MCC. This metric does not prefer any class, because it takes all values from the confusion matrix [46], [47].

H. OPTIMAL CUTOFF VALUE

To show whether adding clinical features changes something in the performance, one ''optimal'' cutoff had to be defined in order to dichotomize the results. Among many approaches for selecting the optimal cutoff, two have similar justifications [48], [49] and both are equally popular. For this work, the Youden index was chosen because the authors of this article agree with Perkins and Schisterman [48], who say that mathematical equations in the medical field should have a rationale in the clinical world. Besides, this index has two desirable properties: a) the independence of the relative size of both groups, and b) all tests that have the same mis-classifications are characterized by the same index value [50].

To sum up, the AUC metric based on sensitivity and specificity was used to assess the discrimination. The MCC was evaluated for each iteration.

IV. RESULTS

As expected, Bayesian networks allowed the discovery of relationships between variables. A comparison of the highest-ranked features of the three models created shows that the strength of clinical features depends on their quality. The more accurate the cytological assessment, the higher its position in the ranking. Figure 5 depicts the strength of clinical features represented by synthetic feature, *clinical risk of malignancy*, estimated by the clinical reference model. Because Bethesda has the most significant impact on clinical risk, rankings were analyzed with attention to this particular variable. The PMBN ranking reveals that *Primary Bethesda* in many samples does not correspond to molecular evidence and can be treated as an observed but confounding variable. In many cases, it is chosen as a second or even third feature. Molecular variables still play the first role. Comparisons of the frequency of particular genes for being chosen in GBN and PMBN suggests that Bethesda only slightly supports molecular data.

A more in-depth analysis of particular samples, presented later in Figure 9, shows that the PMBN has better prediction capabilities in B5 and B6 categories, even though many B5 assessments are mismatched. However, in the dataset used here, mismatched B5 samples were largely underestimated, and samples were malignant; genes were expected to predict malignancy. This corresponds with the results reported in [15], [28] where negative predictive values (NPV)

(Equation 4) for suspicious cytological findings (B5) during the internal evaluation were 85% and even lower in independent ones. NPV is defined as follows:

$$NPV = \frac{TN}{TN + FN} \tag{4}$$

where

TN : the number of true negatives,

FN : is the number of false negatives,

Bethesda was underestimated, but only by about one category in most samples, which still clinically gives a high probability of malignancy. Thus, it improves the PMBN's performance in the B5 category, resulting in fewer FN decisions. The RMBN ranking in Figure 5 depicts that *Clinical Risk* based on *Revised Bethesda* has become the essential feature and significantly changes the proportions of genes playing a key role in malignancy prediction and their importance. *Clinical Risk* was chosen as a node connected with the highest strength to MAL. The *Clinical Risk* is outside of the MB of MAL in only three of the bootstrap tests. This case is fascinating and, because of this, it will be involved in future investigations.

A comparison of the three models' performances shows that the RMBN exceeds the rest of the models, including genes, by about 3-5 percent points. Such a difference may be caused by the clinical data's functional prediction capabilities or a more extensive range of Bethesda categories used in the research. The set of genes selected for analysis were for the samples where cytological assessment was indeterminate (B3 or B4) so during development, B2 and B6 samples were not of interest. In this study, the entire scope of categories was analyzed, hence the performance of molecular classifiers might be lower in border categories.

Figure 6 depicts ROC curves averaged from all bootstraps for all classifiers and evaluated quartiles. Both GBN (b) and PMBN (c) models barely have the same median AUC considering all iterations. There is a difference between the RMBN model (d), and the GBN one (b). The AUC seems to be better for RMBN; the variation is lower, and the median value is higher. The median is significantly better in the clinical classifiers based on *Revised Bethesda* than on *Primary Bethesda*. In both cases, the molecular data provides a solid foundation for a wide range of cutoffs, so almost any threshold value can be chosen without affecting the overall performance.

At first, when the essential clinical variable is mismatched, there is no significant difference between the performance of models built on molecular and clinical features. However, it should be remembered that an imbalanced dataset was used, and that more than 60% of the samples were benign. The optimal threshold for clinical data was about 0.325 of what was necessary to make all B5 and B6 samples malignant. An analysis of the mismatched cytological assessment revealed that most corrections concerned the changes from B5 to B6, which did not affect the RCBN predictions, and from B5 to B3 or B4. That, in turn, made a cut-point shift to the left of the lower value. For this reason, a clinical Bayesian

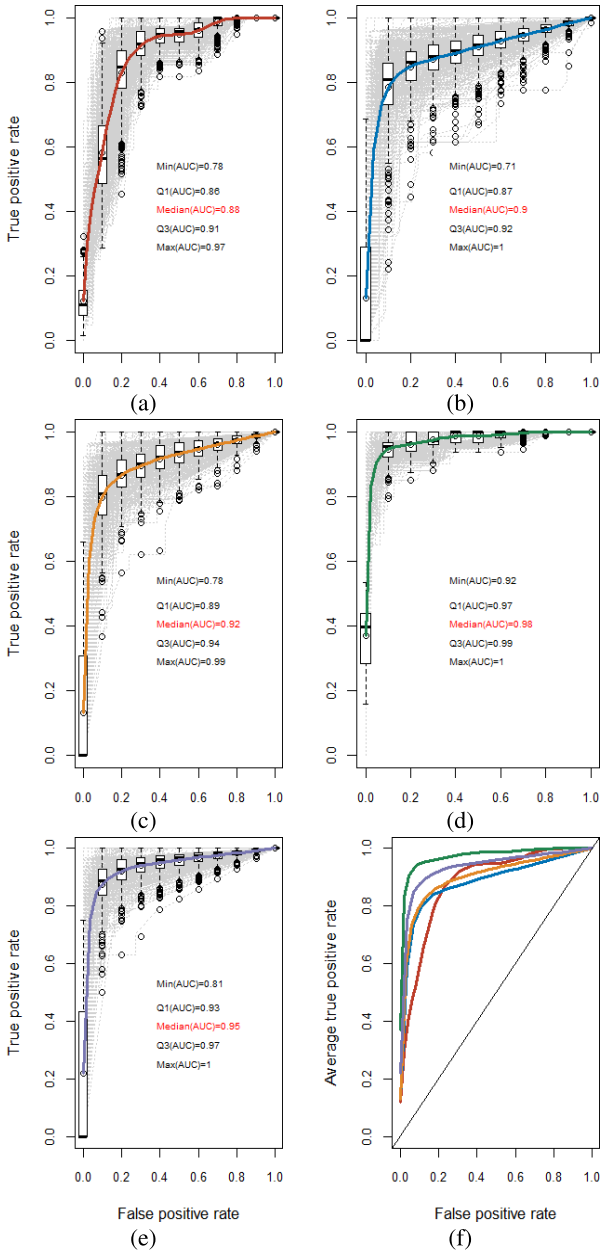


FIGURE 6. ROC curves averaged from all bootstraps and evaluated AUC for (a) clinical data classifier based on Primary Bethesda – PCBN, (b) only molecular data classifier – GBN, (c) molecular data combined with Primary Bethesda – PMBN, (d) clinical data classifier based on Revised Bethesda – RCBN, (e) molecular data combined with Revised Bethesda – RMBN. The last plot presents a comparison of all averaged curves. Whiskers are scaled to get approximately 95% confidence intervals.

classifier based on *Revised Bethesda* had such a high AUC even though the optimal cutoff was lower.

The impact of genes can be seen in the middle of the false-positive rate data. It should be taken into account that a gene set was used that was developed to support the prediction of benign lesions (rule-out molecular classifier) for which performance was initially evaluated in the range of B3 - B4 and for some B5 samples. Where the clinical classifier based on *Primary Bethesda* stayed the same, the

molecular classifier slowly increased its ability to rank randomly chosen positive samples higher than randomly chosen negative ones after a certain cut-point. The reason for this was because of the better prediction of indeterminate samples (B3 or B4). When integrated into one model, they became complementary. It was also noticed that RCBN (clinical classifier based on *Revised Bethesda*) outperforms RMBN (combined one). Concerning the authors' initial questions, it has just been shown that it is better to combine clinical and molecular features in one model than to integrate the results of the molecular analysis into a clinical context later on.

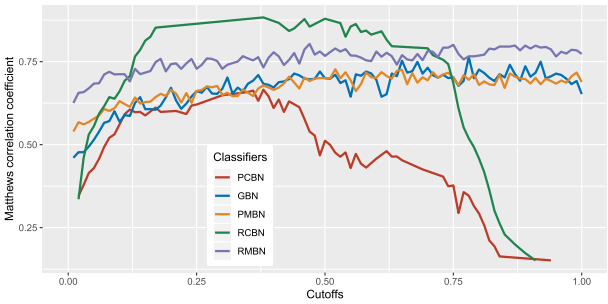


FIGURE 7. Matthews correlation coefficients for all cutoffs. PCBN [Mean – 0.48, Median – 0.48], GBN [0.67, 0.69], PMBN [0.67,0.68], RCBN [0.65, 0.74], RMBN [0.76, 0.76]. Colors correspond to the colors in Figure 6.

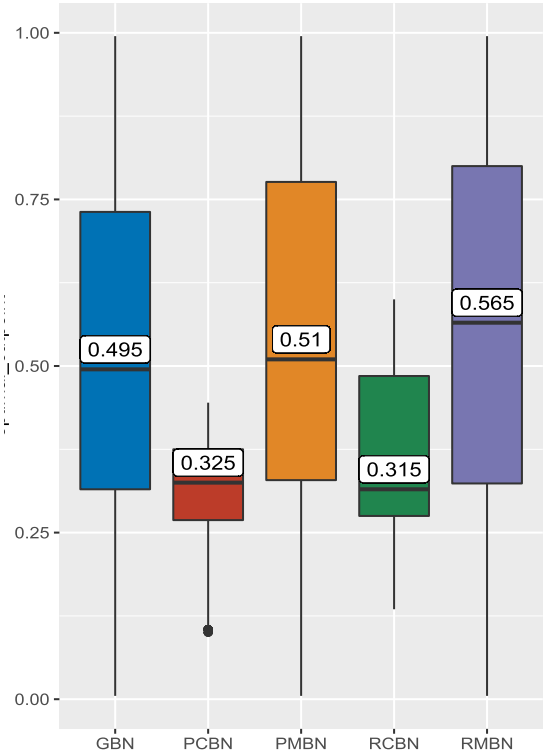


FIGURE 8. Younden's index – median value from all bootstraps.

Having regard to the results of the experiments with imbalanced dataset reported in [46], [51], we decided to compare our classifiers using the MCC metric to handle data imbalance; the disproportion ratio was 1.64. Figure 7 depicts the comparison of averaged MCC values for all the developed classifiers and cutoffs. It is worth mentioning that a combination of clinical and molecular features reduces variations among all the possible cutoffs and brings mean and median close to each other. As a result, choosing "the best cut-point" does not affect the final performance as much. With regard to PCBN and RCBN (clinical classifiers), the overall quality significantly depends on the researcher's choice of cut-point. The green line represents MCC for RCBN (clinical classifier based on *Revised Bethesda*), and the purple one describes MCC for RMBN, when the same variable is integrated with molecular data. In both types of models, the quality of binary classification increased significantly.

In further analyzes, results achieved for a specific cutoff chosen based on the median of Younden's index was reported, as presented in Figure 8.

The impact of clinical data quality on performance is also visible when increasing the value of optimal cut-point between PMBN and RMBN. The cutoff is about 6 points higher, which mainly affects the proper prediction of B5 samples. For samples where Bethesda is B5 or B6, and genes are not clear for the benign tumor, predictions are mainly malignant (B5 and B6 clinical results show evidence of malignancy). Samples with strong gene evidence are benign.

Figure 9 shows a detailed comparison of each model's prediction capabilities for all bootstraps: the left side shows samples with cytological assessment B5; colors correspond to the results of a classification in every bootstrap. The revised category for each sample was put in the middle. The right side presents models that rely on those revised values.

The difference between PCBN and RCBN is visible because the clinical models are data- (sample-) independent. They are built on external data from the clinical database. In every bootstrap, they predict the same value. The only parameter which can have an impact on the performance is the cut-point. Figure 9 shows that those samples which were cytologically overestimated are malignant according to PCBN but are benign in the majority according to RCBN.

The molecular and combined models are sample sensitive, but when they rely on features from different sources, those features become complementary. Genes give better results when the sample is benign. *Clinical Risk*, in turn, reduces the number of False Negative instances when the sample is malignant; when it is based on *Revised Bethesda*, it reduces further. It must be remembered that the sample size was small, and there is no sharp border of probability between adjacent categories, apart from those at the border. Therefore, the use of machine learning methods (mainly unsupervised) on "as is" datasets may lead to wrong decisions when new data is introduced.

V. DISCUSSION

The overall average MCC for various cutoff values shows that molecular features provide a solid foundation. Even mismatched Bethesda categories cannot significantly reduce

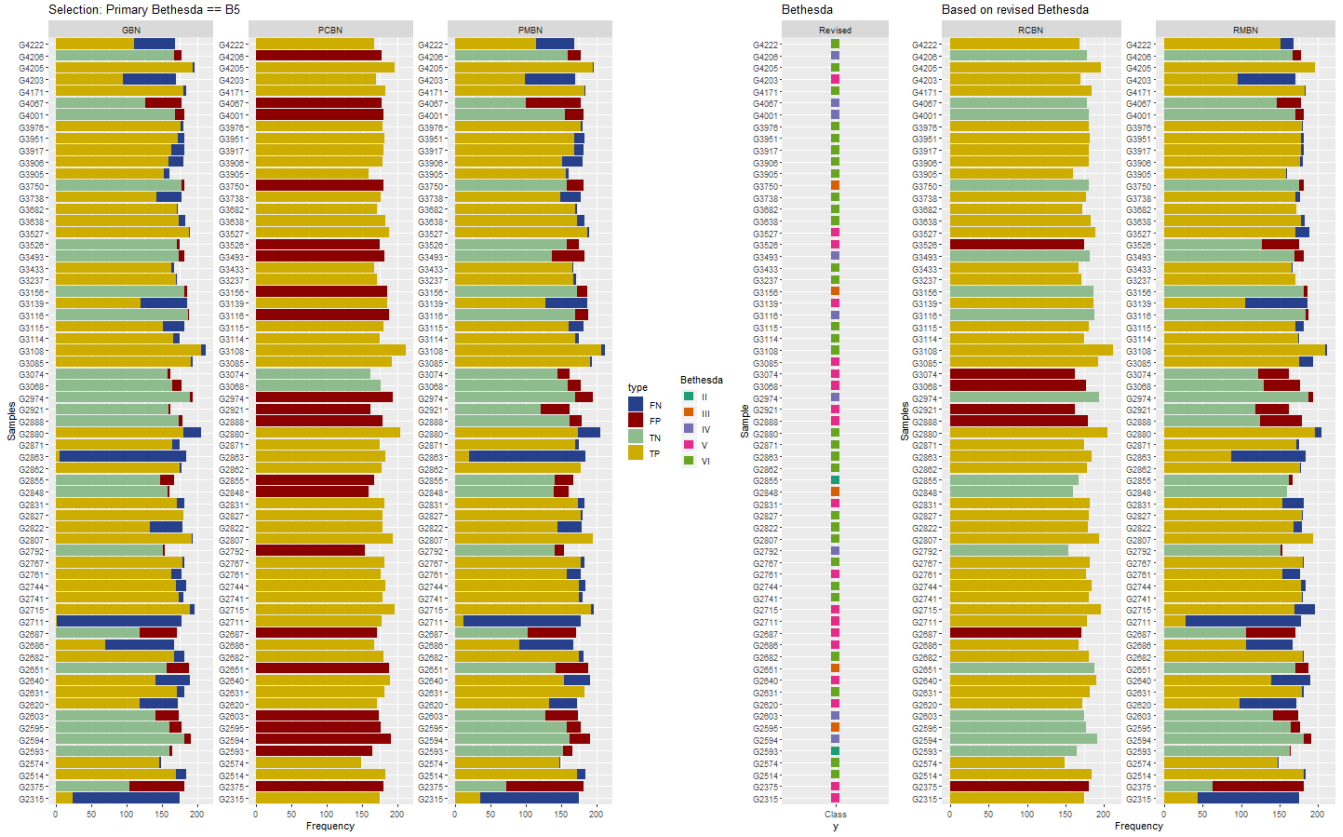


FIGURE 9. Comparison of prediction capability of developed classifiers in all bootstraps. Plot with *Revised* title depicts the Bethesda category after tissue revision by experienced pathologists from NCI. The cutoffs evaluated after averaging the ROC curves are as follows: 0.495/0.325/0.51/0.315/0.565. The figure presents samples initially assessed as B5 (suspicious for malignancy).

the accuracy of classifiers. The strength of the molecular set is visible when compared to the clinical ones where accuracy decreases drastically for higher cutoffs. However, as presented in this article, adding synthetic clinical features based on Bethesda to the set of initial features and the usage of Bayesian networks, enables strong dependencies to be discovered between them and the molecular data. The strength depends on the quality of Bethesda as evidence, and this can vary from center to center [52], [53]. The difference in specialists' experience and the role of the center in TC treatment explains the reason for this existing variety.

Please note, that in Figure2, the risk of malignancy of Polish patients in category B3 could be underestimated due to the fact that patients diagnosed in NCI are referred to other centers for surgery and may not be in the observations. The authors are aware of this, but in their opinion the higher risk of malignancy should only improve the results of the combined data models.

Bayesian networks allow researchers and medics to update their beliefs even if it is only partial, hard evidence or suspicion (soft evidence) that exists. This may help in the daily practice of pathologists who can query the network to get the most probable values of particular genes or assess the probability of co-existence.

In the diagnostic process of the thyroid nodule, the first step is to analyze clinical features. The decision about whether to use the molecular test is made based on them. Currently, Bethesda is used as a selection criterion in order to check whether the patient should be tested. If the initial assessment of the Bethesda categorization is undervalued (e.g. B3 when it is actually B5), that may result in misclassification due to a much lower specificity, i.e. false negatives. On the other hand, when the score is overestimated (i.e. B5 is in fact, B3), the patient may be deprived of the opportunity to be tested. Proper interpretation of all these test results requires integrating them [4] in the clinical context, meaning that decisions about surgery should incorporate the consistency of both clinical and molecular results.

Many papers describe the impact of clinical features on the patient management process [54], [55], and some researches have started taking into account the combination of clinical and molecular data in the management of the thyroid nodule in the area of molecular test development [56]. However, no studies focusing on the quality of clinical data and its impact on analysis have been reported. The use of unsupervised methods in mining patterns in data is vast, but in the context of a particular diagnostic process, one should be aware of the quality impact on test development. So far, molecular tests that have been introduced are aimed to support physicians when indeterminate nodules are found, but what if the clinical assessment is incorrect? In this article, it was confirmed that based on molecular data, benign

samples could be predicted well, but the prediction capabilities are better when combined with clinical data. Even if Bethesda is mismatched, the performance does not deteriorate. Available independent research of molecular tests in TC highlights the necessity of integrating molecular tests results into a clinical context. Thus, the context should be reliable.

In future research, the same Bayesian networks can be used to detect whether the provided Bethesda category could be mismatched. Bethesda is a factor variable, but it determines a risk of malignancy, which in turn is measured on a continuous scale within clinical practice. Because of overlapping values (different categories may cover the same risk), there is a problem with conflict analysis between clinical and molecular evidence and identification of true negative correlation. On the other hand, BNs learned from combined data sources, capture the probability distribution over all the nodes of interest (both molecular and clinical), which could be explored towards the identification of the true Bethesda's value. The conditional queries using the Bethesda's Markov blanket is the right direction. Also, the authors want to investigate why the *Clinical Risk* was excluded from the MB of MAL for three iterations. Last but not least is the problem of the proper method of discretization. This could affect the performance of each model.

ACKNOWLEDGMENT

The ethics committee approved full protocol of the study.

REFERENCES

[1] L. M. Lowenstein, S. P. Basourakos, M. D. Williams, P. Troncoso, J. R. Gregg, T. C. Thompson, and J. Kim, "Active surveillance for prostate and thyroid cancers: Evolution in clinical paradigms and lessons learned," *Nature Rev. Clin. Oncol.*, vol. 16, no. 3, pp. 168–184, Mar. 2019. [Online]. Available: http://www.nature.com/articles/s41571-018-0116-x

[2] D. Rusinek, E. Chmielik, J. Krajewska, M. Jarzab, M. Oczko-Wojciechowska, A. Czarniecka, and B. Jarzab, "Current advances in thyroid cancer Management. Are we ready for the epidemic rise of diagnoses?" *Int. J. Mol. Sci.*, vol. 18, no. 8, p. 1817, Aug. 2017. [Online]. Available: http://www.mdpi.com/1422-0067/18/8/1817

[3] M. Oczko-Wojciechowska, A. Kotecka-Blicharz, J. Krajewska, D. Rusinek, M. Barczyński, B. Jarząb, and A. Czarniecka, "European perspective on the use of molecular tests in the diagnosis and therapy of thyroid neoplasms," *Gland Surgery*, vol. 9, no. S2, pp. S69–S76, Feb. 2020. [Online]. Available: http://gs.amegroups.com/article/view/31498/28202

[4] R. Jug and X. Jiang. *Molecular Testing in FNA*. Accessed: May 10, 2020. [Online]. Available: https://www.pathologyoutlines.com/topic/thyroidglandmolectestingfna.html

[5] B. R. Haugen, E. K. Alexander, K. C. Bible, G. M. Doherty, S. J. Mandel, Y. E. Nikiforov, F. Pacini, G. W. Randolph, A. M. Sawka, M. Schlumberger, K. G. Schuff, S. I. Sherman, J. A. Sosa, D. L. Steward, R. M. Tuttle, and L. Wartofsky, "2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer," *Thyroid*, vol. 26, no. 1, pp. 1–133, Jan. 2016. [Online]. Available: https://www.liebertpub.com/doi/10.1089/thy.2015.0020

[6] B. Yet, Z. B. Perkins, T. E. Rasmussen, N. R. M. Tai, and D. W. R. Marsh, "Combining data and meta-analysis to build Bayesian networks for clinical decision support," *J. Biomed. Informat.*, vol. 52, pp. 373–385, Dec. 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046414001816

[7] D. Zhao and C. Weng, "Combining PubMed knowledge and EHR data to develop a weighted Bayesian network for pancreatic cancer prediction," *J. Biomed. Informat.*, vol. 44, no. 5, pp. 859–868, Oct. 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046411000955

[8] Z. Liao, D. Li, X. Wang, L. Li, and Q. Zou, "Cancer diagnosis through IsomiR expression with machine learning method," *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, Feb. 2018.

[9] Q. Zou and Q. Ma, "The application of machine learning to disease diagnosis and treatment," *Math. Biosciences*, vol. 320, Feb. 2020, Art. no. 108305. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0025556419305462

[10] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-Resp-forest: A deep forest model to predict anti-cancer drug response," *Methods*, vol. 166, pp. 91–102, Aug. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1046202318303232

[11] M. A. J. van Gerven, B. G. Taal, and P. J. F. Lucas, "Dynamic Bayesian networks as prognostic models for clinical patient management," *J. Biomed. Informat.*, vol. 41, no. 4, pp. 515–529, Aug. 2008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046408000154

[12] N. Fenton and M. Neil, "Comparing risks of alternative medical diagnosis using Bayesian arguments," *J. Biomed. Informat.*, vol. 43, no. 4, pp. 485–495, Aug. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046410000195

[13] J. Yu, Z. Xuan, X. Feng, Q. Zou, and L. Wang, "A novel collaborative filtering model for LncRNA-disease association prediction based on the Naïve Bayesian classifier," *BMC Bioinf.*, vol. 20, no. 1, p. 396, Dec. 2019.

[14] Y. E. Nikiforov, S. E. Carty, S. I. Chiosea, C. Coyne, U. Duvvuri, R. L. Ferris, W. E. Gooding, S. O. LeBeau, N. P. Ohori, R. R. Seethala, M. E. Tublin, L. Yip, and M. N. Nikiforova, "Impact of the multi-gene ThyroSeq next-generation sequencing assay on cancer diagnosis in thyroid nodules with atypia of undetermined Significance/Follicular lesion of undetermined significance cytology," *Thyroid*, vol. 25, no. 11, pp. 1217–1223, Nov. 2015. [Online]. Available: https://www.liebertpub.com/doi/10.1089/thy.2015.0305

[15] E. K. Alexander, G. C. Kennedy, Z. W. Baloch, E. S. Cibas, D. Chudova, J. Diggans, L. Friedman, R. T. Kloos, V. A. LiVolsi, S. J. Mandel, S. S. Raab, J. Rosai, D. L. Steward, P. S. Walsh, J. I. Wilde, M. A. Zeiger, R. B. Lanman, and B. R. Haugen, "Preoperative diagnosis of benign thyroid nodules with indeterminate cytology," *New England J. Med.*, vol. 367, no. 8, pp. 705–715, 2012. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMoa1203208

[16] H. E. González *et al.*, "A 10-gene classifier for indeterminate thyroid nodules: Development and multicenter accuracy study," *Thyroid*, vol. 27, no. 8, pp. 1058–1067, Aug. 2017. [Online]. Available: https://www.liebertpub.com/doi/10.1089/thy.2017.0067

[17] M. T. D. Santos, A. L. Buzolin, R. R. Gama, E. C. A. D. Silva, R. M. Dufloth, D. L. A. Figueiredo, and A. L. Carvalho, "Molecular classification of thyroid nodules with indeterminate cytology: Development and validation of a highly sensitive and specific new miRNA-based classifier test using fine-needle aspiration smear slides," *Thyroid*, vol. 28, no. 12, pp. 1618–1626, Dec. 2018. [Online]. Available: https://www.liebertpub.com/doi/10.1089/thy.2018.0254

[18] B. Jarząb *et al.*, "Guidelines of Polish National Societies diagnostics and treatment of thyroid carcinoma. 2018 update," *Endokrynologia Polska*, vol. 69, no. 1, pp. 34–74, 2018.

[19] F. N. Tessler, W. D. Middleton, E. G. Grant, J. K. Hoang, L. L. Berland, S. A. Teefey, J. J. Cronan, M. D. Beland, T. S. Desser, M. C. Frates, L. W. Hammers, U. M. Hamper, J. E. Langer, C. C. Reading, L. M. Scoutt, and A. T. Stavros, "ACR thyroid imaging, reporting and data system (TI-RADS): White paper of the ACR TI-RADS committee," *J. Amer. College Radiol.*, vol. 14, no. 5, pp. 587–595, May 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1546144017301862

[20] S. Mondal, S. Sinha, B. Basak, D. Roy, and S. Sinha, "The bethesda system for reporting thyroid fine needle aspirates: A cytologic study with histologic follow-up," *J. Cytology*, vol. 30, no. 2, p. 94, 2013. [Online]. Available: http://www.jcytol.org/text.asp?2013/30/2/94/112650

[21] S. Alshaikh, Z. Harb, E. Aljufairi, and S. A. Almahari, "Classification of thyroid fine-needle aspiration cytology into Bethesda categories: An institutional experience and review of the literature," *CytoJournal*, vol. 15, p. 4, Feb. 2018. [Online]. Available: https://doi.org/10.4103/cytojournal.cytojournal_32_17

[22] R. M. Tuttle, B. Haugen, and N. D. Perrier, "Updated American joint committee on Cancer/Tumor-Node-Metastasis staging system for differentiated and anaplastic thyroid cancer (Eighth Edition): What changed and why?" *Thyroid*, vol. 27, no. 6, pp. 751–756, Jun. 2017. [Online]. Available: https://www.liebertpub.com/doi/10.1089/thy.2017.0102

[23] V. Zaydfudim, I. D. Feurer, M. R. Griffin, and J. E. Phay, "The impact of lymph node involvement on survival in patients with papillary and follicular thyroid carcinoma," *Surgery*, vol. 144, no. 6, pp. 1070–1078, Dec. 2008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0039606008005527
[24] I. J. Nixon *et al.*, "An international multi-institutional validation of age 55 years as a cutoff for risk stratification in the AJCC/UICC staging system for well-differentiated thyroid cancer," *Thyroid*, vol. 26, no. 3, pp. 373–380, Mar. 2016. [Online]. Available: https://www.liebertpub.com/doi/10.1089/thy.2015.0315
[25] A. Kelly, B. Barres, F. Kwiatkowski, M. Batisse-Lignier, B. Aubert, C. Valla, F. Somda, F. Cachin, I. Tauveron, and S. Maqdasy, "Age, thyroglobulin levels and ATA risk stratification predict 10-year survival rate of differentiated thyroid cancer patients," *PLoS ONE*, vol. 14, no. 8, Aug. 2019, Art. no. e0221298. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0221298
[26] J. Lortet-Tieulent, S. Franceschi, L. Dal Maso, and S. Vaccarella, "Thyroid cancer 'epidemic' also occurs in low- and middle-income countries," *Int. J. Cancer*, vol. 144, no. 9, pp. 2082–2087, May 2019, doi: 10.1002/ijc.31884.
[27] E. Goodarzi, A. Moslem, H. Feizhadad, A. Jarrahi, H. Adineh, M. Sohrabivafa, and Z. Khazaei, "Epidemiology, incidence and mortality of thyroid cancer and their relationship with the human development index in the world: an ecology study in 2018," *Adv. Hum. Biol.*, vol. 9, no. 2, p. 162, 2019.
[28] C. Ferraz, "Can current molecular tests help in the diagnosis of indeterminate thyroid nodule FNAB?" *Arch. Endocrinology Metabolism*, vol. 62, no. 6, pp. 576–584, 2018. [Online]. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2359-39972018000600576&lng=en&nrm=iso
[29] D. Chudova, J. I. Wilde, E. T. Wang, H. Wang, N. Rabbee, C. M. Egidio, J. Reynolds, E. Tom, M. Pagan, C. T. Rigl, L. Friedman, C. C. Wang, R. B. Lanman, M. Zeiger, E. Kebebew, J. Rosai, G. Fellegara, V. A. LiVolsi, and G. C. Kennedy, "Molecular classification of thyroid nodules using high-dimensionality genomic data," *J. Clin. Endocrinol. Metabolism*, vol. 95, no. 12, pp. 5296–5304, Dec. 2010. [Online]. Available: https://academic.oup.com/jcem/article-lookup/doi/10.1210/jc.2010-1087
[30] J. Pearl, TotalBoox, and TBX,*Probabilistic Reasoning in Intelligent Systems*. Amsterdam, The Netherlands: Elsevier, 2014. [Online]. Available: http://www.totalboox.com/book/id-4660707859007505083
[31] F. V. Jensen, *Bayesian Networks and Decision Graphs*. New York, NY, USA: Springer, 2001. [Online]. Available: http://link.springer.com/10.1007/978-1-4757-3502-4
[32] R. Nagarajan, M. Scutari, and S. Lèbre, *Bayesian Networks in R*. New York, NY, USA: Springer, 2013. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-6446-4
[33] N. Friedman, M. Goldszmidt, and A. Wyner, "Data analysis with Bayesian networks: A bootstrap approach," 2013, *arXiv:1301.6695*. [Online]. Available: http://arxiv.org/abs/1301.6695
[34] P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, and M. J. Druzdzel, "Bayesian networks for risk prediction using real-world data: A tool for precision medicine," *Value Health*, vol. 22, no. 4, pp. 439–445, Apr. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1098301519300579
[35] P. Kaewprag, C. Newton, B. Vermillion, S. Hyun, K. Huang, and R. Machiraju, "Predictive models for pressure ulcers from intensive care unit electronic health records using Bayesian networks," *BMC Med. Informat. Decis. Making*, vol. 17, no. S2, p. 65, Jul. 2017. [Online]. Available: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0471-z
[36] A. Placzek, A. Płuciennik, M. Pach, M. Jarząb, and D. Mrozek, "The role of feature selection in text mining in the process of discovering missing clinical annotations–Case study," in *Beyond Databases, Architectures and Structures. Paving the Road to Smart Data Processing and Analysis*, vol. 1018, S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, and D. Kostrzewa, Eds. Cham, Switzerland: Springer, pp. 248–262. [Online]. Available: http://link.springer.com/10.1007/978-3-030-19093-4_19
[37] H. Bengtsson, K. Simpson, J. Bullard, and K. M. Hansen, "Aroma. affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory," Dept. Statist., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 745, Feb. 2008.
[38] *Microarray Lab*. Accessed: Apr. 16, 2020. [Online]. Available: http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/
[39] R. R. Seethala, Z. W. Baloch, J. A. Barletta, E. Khanafshar, O. Mete, P. M. Sadow, V. A. LiVolsi, Y. E. Nikiforov, G. Tallini, and L. D. Thompson, "Noninvasive follicular thyroid neoplasm with papillary-like nuclear features: A review for pathologists," *Mod. Pathol.*, vol. 31, no. 1, pp. 39–55, 2018. [Online]. Available: http://www.nature.com/articles/modpathol2017130
[40] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007. [Online]. Available: http://science.sciencemag.org/content/315/5814/972.abstract
[41] P. Thavikulwat, "Affinity propagation: A clustering algorithm for computer-assisted business simulations and experiential exercises," in *Proc. Annu. ABSEL Conf.*, vol. 35, 2014, pp. 220–224. [Online]. Available: https://absel-ojs-ttu.tdl.org/absel/index.php/absel/article/view/408
[42] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Model. Methods Appl. Sci.*, vol. 1, no. 4, pp. 300–307, 2007.
[43] P. Mahanta, H. A. Ahmed, J. K. Kalita, and D. K. Bhattacharyya, "Discretization in gene expression data analysis: A selected survey," in *Proc. 2nd Int. Conf. Comput. Sci., Eng. Inf. Technol. (CCSEIT)*, Oct. 2012, pp. 69–75. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2393216.2393229
[44] C. A. Gallo, R. L. Cecchini, J. A. Carballido, S. Micheletto, and I. Ponzoni, "Discretization of gene expression data revised," *Briefings Bioinf.*, vol. 17, no. 5, pp. 758–770, 2016. [Online]. Available: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv074
[45] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Mach. Learn.*, vol. 77, no. 1, pp. 103–123, Oct. 2009. [Online]. Available: http://link.springer.com/10.1007/s10994-009-5119-5
[46] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020. [Online]. Available: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7
[47] J. C. D. Lopes, F. M. dos Santos, A. Martins-José, K. Augustyns, and H. De Winter, "The power metric: A new statistically robust enrichment-type metric for virtual screening applications with early recovery capability," *J. Cheminformatics*, vol. 9, no. 1, Dec. 2017. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-016-0189-4
[48] N. J. Perkins and E. F. Schisterman, "The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve," *Amer. J. Epidemiol.*, vol. 163, no. 7, pp. 670–675, 2006. [Online]. Available: http://academic.oup.com/aje/article/163/7/670/77813/The-Inconsistency-of-Optimal-Cutpoints-Obtained
[49] I. Unal, "Defining an optimal cut-point value in ROC analysis: An alternative approach," *Comput. Math. Methods Med.*, vol. 2017, May 2017, Art. no. 3762651. [Online]. Available: https://www.hindawi.com/journals/cmmm/2017/3762651/
[50] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
[51] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using matthews correlation coefficient metric," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0177678. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0177678
[52] A. Crowe, A. Linder, O. Hameed, C. Salih, J. Roberson, J. Gidley, and I. A. Eltoum, "The impact of implementation of the Bethesda System for Reporting Thyroid Cytopathology on the quality of reporting, 'risk' of malignancy, surgical rate, and rate of frozen sections requested for thyroid lesions," *Cancer Cytopathol.*, vol. 119, no. 5, pp. 315–321, Oct. 2011. [Online]. Available: http://doi.wiley.com/10.1002/cncy.20174
[53] S. Kannan, N. Raju, V. Kekatpure, N. Chandrasekhar, V. Pillai, A. Keshavamurthy, M. Kuriakose, P. Rekha, N. Raghavan, A. Lakhsmikantha, S. Ramaiah, and B. Dave, "Improving bethesda reporting in thyroid cytology: A team effort goes a long way and still miles to go," *Indian J. Endocrinol. Metabolism*, vol. 21, no. 2, p. 277–281, 2017. [Online]. Available: http://www.ijem.in/text.asp?2017/21/2/277/202032
[54] W. C. Faquin, R. Izquierdo, and K. K. Khurana, "FNA of misclassified primary malignant neoplasms of the thyroid: Impact on clinical management," *CytoJournal*, vol. 6, p. 1, Jan. 2009. [Online]. Available: https://cytojournal.com/fna-of-misclassified-primary-malignant-neoplasms-of-the-thyroid-impact-on-clinical-management/
[55] A. W. Phillips, J. D. Fenwick, U. K. Mallick, and P. Perros, "The impact of clinical guidelines on surgical management in patients with thyroid cancer," *Clin. Oncol.*, vol. 15, no. 8, pp. 485–489, Dec. 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S093665550300195X

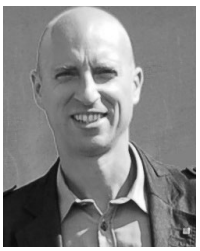
[56] H. Lasolle, B. Riche, M. Decaussin-Petrucci, E. Dantony, V. Lapras, C. Cornu, J. Lachuer, J.-L. Peix, J.-C. Lifante, O.-M. Capraru, S. Selmi-Ruby, B. Rousset, F. Borson-Chazot, and P. Roy, "Predicting thyroid nodule malignancy at several prevalence values with a combined Bethesda-molecular test," *Transl. Res.*, vol. 188, pp. 58–66, Oct. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1931524417301421

ALEKSANDER PŁACZEK received the M.Sc. degree in computer science from the Silesian University of Technology, Gliwice, Poland, in 2000, where he is currently pursuing the Ph.D. degree. He is a member of the Implementation Doctorate Program supported by the Polish Ministry of Science. During the last eight years, he held the position of the Deputy Manager of the Research and Development Department, WASKO S. A., Poland. At the same time, he acted as a Software Architect in many innovative projects conducted in cooperation with Polish reference centers: National Cancer Institute and the Silesian Centre for Heart Diseases. He has been involved in many research projects from medical and biotechnological areas such as MILESTONE—Molecular diagnostics and imaging in individualized therapy for breast, thyroid and prostate cancer, BIOTEST—Remote platform for hypothesis testing and analysis of biomedical data, and MONITEL—HF: The use of teletransmission of medical data in patients with heart failure for improvement of quality of life and reduction of treatment costs. The results of those studies were published as conference papers and implemented as software products. Since the beginning of May 2020, he holds the position of the Project Director of the Department of Analytics, Research and Development, GABOS Software, and the WASKO Capital Group, Poland. His research interests include exploring the medical databases and learning Bayesian networks from different data sources, conflict, and sensitive analysis. He has been working on oncological diagnostics based on omic data supported by the clinical picture of patients with a similar disease profile.

ALICJA PŁUCIENNIK received the M.Sc. degree in biotechnology (bioinformatics) from the Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland, in 2017. She is currently pursuing the Ph.D. degree with the Department of System Biology and Engineering, Silesian University of Technology, in fields of biomedical engineering and computer science. She has been working as a Telemedicine Systems Analyst with WASKO S. A., Gliwice, since 2018. Her research interests include improvements in cancer diagnostics, data integration and molecular markers selection, and designing researcher friendly systems. From 2016 to 2017, she received the Scholarship of the Polish National Science Center (NSC) for research on evolution of gating amino acids in DynaGate project carried out by the Tunneling Group.

AGNIESZKA KOTECKA-BLICHARZ graduated from the Harvard Medical School Course in clinical endocrinology, Boston, USA. She received the M.D. degree from Silesian Medical University, in 2000. She is currently an Endocrinologist with the Department of Nuclear Medicine and Endocrine Oncology, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Poland. She is also a Clinical practitioner in the field of endocrine oncology. She received the title of a Specialist in endocrinology in 2013. Her research interests include rare endocrine neoplasm and its genetic background. She is also focused on molecular diagnostics and imaging of thyroid neoplasm under the Project MILESTONE—Molecular diagnostics and imaging in individualized therapy for breast, thyroid, and prostate cancer. She is a coauthor of Polish National Societies Diagnostics and Treatment of Thyroid Carcinoma Guidelines. She is a member of the Polish Endocrine Society, the European Society of Endocrinology, and the Endocrine Society.

MICHAŁ JARZĄB received the M.D. and Ph.D. degrees. He graduated from Silesian Medical University, in 2001. He holds a position of the Head of the Center of Diagnostics and Therapy of Breast Cancer (Breast Unit), Maria Skłodowska-Curie National Research Institute of Oncology, Gliwice, Poland. His Ph.D. thesis was devoted to the expression of membrane iodine transporters in breast cancer. Previously, he was appointed at the Department of Tumor Biology, the Center of Translational Research, the Department of Clinical and Experimental Oncology. Since 2011, he has been with the IIIrd Department of Radiotherapy and Chemotherapy. He is also a Board-Approved Specialist in clinical oncology, was involved as a clinical investigator in numerous clinical trials. He is the author or a coauthor of more than 60 original publications, with Hirsch index 16. Since 2001, he has been involved in the development and research activity of one of the first Polish laboratory implementing methods of functional genomics into translation oncology. His research interests include translational genomics of cancer, basic research with the use of transcriptomic techniques and clinical usefulness of classic and molecular predictive factors. He is a coauthor of Polish diagnostic and therapeutic guidelines in oncology, a member of multiple societies, including the American Association of Clinical Oncology, the European Society of Medical Oncology, and the European Society for Radiotherapy and Oncology. He is a member of Board of the Polish Society of Clinical Oncology and the National Societies Committee of European Society of Medical Oncology.



DARIUSZ MROZEK (Senior Member, IEEE) received the Ph.D. degree from SUT, in 2006. He is currently an Associate Professor and the Head of the Department of Applied Informatics, Silesian University of Technology (SUT), Gliwice, Poland. His research interests include the IoT, parallel and cloud computing, databases, big data, and bioinformatics. His current research interests include developing the IoT solutions for healthcare on the use of novel computation techniques to get insights from biological data, including NGS and proteomics data. He is the author of more than 90 articles published in conference proceedings and international journals, and two books on the use of big data analytics and high-performance computations in protein bioinformatics published by Springer; a Co-Editor of fifteen other books devoted to databases and data processing; and an Editor of many special issues in reputable scientific journals. He is a member of the IEEE Engineering in Medicine and Biology Society (EMBS), the IEEE Systems, Man, and Cybernetics Society (SMCS), and IEEE Cloud Computing Community. He has collaborated with qualified institutions by working in different research projects, like the Imperial College of London, U.K., Amazon, USA, or Microsoft Research, USA.