MIAT: A Novel Attribute Selection Approach to Better Predict Upper Gastrointestinal Cancer

Avi Rosenfeld\textsuperscript{1}, David G. Graham\textsuperscript{2}, Rifat Hamoudi\textsuperscript{3}, Rommel Butawan\textsuperscript{2}, Victor Eneh\textsuperscript{2}, Saif Khan\textsuperscript{3}

\textsuperscript{1}Department of Industrial Engineering, Jerusalem College of Technology (JCT), Israel
\textsuperscript{2}Department of Gastroenterology, University College London Hospital (UCLH), United Kingdom
\textsuperscript{3}Division of Surgery and Interventional Science, University College London, United Kingdom

\begin{abstract}
The use of data mining has led to many significant medical discoveries. However, many challenges still exist in using these methods for knowledge discovery within this field given that the large amounts of data medical practitioners collect often creates a curse of dimensionality. To address this challenge, attribute selection approaches have been developed. However, current approaches typically put equal weight on all values within that attribute. At times, and especially within medical domains, we claim that these approaches might miss attributes where only a small subset of attribute values contain a strong indication for one of the target values and thus should still be selected.

To quantify this approach, we present MIAT, an algorithm that defines Minority Interesting Attribute Thresholds to find these important attribute values. As we developed MIAT to help better diagnose upper gastrointestinal cancer, we present how we use the attributes selected through this approach to build a predictive model for this cancer. To demonstrate MIAT’s generality, we also applied it to a canonical Hungarian Heart Disease Dataset. In both datasets we found that MIAT yields significantly better accuracy and sensitivity over traditional attribute selection approaches.
\end{abstract}

I. INTRODUCTION

Upper gastrointestinal cancer is a leading cause of death worldwide particularly due to its combined approximate 5-year survival of only 17\textsuperscript{\%}\textsuperscript{1}. It is also becoming more prevalent \cite{1}. Consequently, vast resources have been allocated to help detect and prevent such cancers. The existing system in the UK relies upon patients’ primary care physicians to refer them for specialist investigation based on clinical suspicion, known as the “Two-week wait” expedited referral pathway. Unfortunately this process is inefficient. For example, 86-97\% of expedited referrals for suspected gastrointestinal cancers do not yield this diagnosis \cite{2, 3}. Equally important, current practices have produced a negligible improvement in survival \cite{1} as they often fail to identify these cancers in their early or high risk pre-cancerous stages. Unfortunately, once the onset of symptoms and signs has occurred the patient often has advanced, incurable disease \cite{3}. However, if caught early or in the pre-cancerous stage, these lesions can be easily removed through minimally invasive, endoscopic techniques \cite{6, 8}. Therefore, the early detection of people at risk for having these pre-cancerous lesions is critical.

Existing monitoring programs rely on established surveillance programs and the expedited referral pathway (“Two-week wait”) for suspected upper gastrointestinal (UGI) cancer. Published data demonstrates that these surveillance programs are ineffective \cite{9, 10} and the “Two-week wait” system has long been criticized for demonstrating no survival benefit for patients \cite{11}. Thus, the goal of our research has been to determine if one could better identify the at-risk individual based on easily obtainable information with excellent sensitivity (recall of cancer patients) while still achieving high accuracy and fair specificity.

Towards achieving this goal, we have embarked on an ambitious program of data mining peoples’ demographic and medical data to try to ascertain which factors, if any, can help predict a person’s risk of developing UGI cancer. Data mining medical data is not new, and models have typically focused on using a set of patient-specific information to predict a medical outcome or to help support doctors make a clinical diagnosis \cite{12}. Several industrial partners are also contributing to this field. One example is IBM’s WatsonPath project in conjunction with the Cleveland Clinic\textsuperscript{2}.

Despite advances, the medical field has lagged behind in effectively applying machine learning. This paper addresses one particular challenge – of how to accurately identify which attributes are important for learning the desired target variable – here whom will develop cancer. This task is important to tame the phenomenon known as the “curse of dimensionality” \cite{13, 14} whereby having too many attributes creates state space too large for existing algorithms to build accurate models. To date, many such algorithms have been proposed which could generally be categorized into attribute filtering, wrappers and embedded algorithms. A survey of these approaches is presented in the next section.

This paper’s key contribution is MIAT, an algorithm that defines Minority Interesting Attribute Thresholds. MIAT focuses on identifying important features where only the minority of an attribute’s values strongly point to one of the classes needing to be learned. We claim that at times it can be important to select such attributes, especially within the medical domain we considered. However, attribute selection approaches to date typically treat all values within a given

\begin{flushleft}
\textsuperscript{1}info.cancerresearchuk.org/cancerstats/, http://globocan.iarc.fr/
\textsuperscript{2}https://www.research.ibm.com/cognitive-computing/watson/watsonpaths.shtml
\end{flushleft}
attribute equally, and thus focus on the general importance of all values within a given attribute, or combinations of the full set of different attributes’ values [15], [16]. As such, these approaches would typically not select attributes based on strong indications within a minority subset of values.

To make this general idea clearer, consider the following example. Assume that people can be categorized into different groups based on how much they smoke – non-smoker, occasional smoker, daily smoker, heavy daily smoker. Assume a population of 1000 where only 5% of considered population can be categorized as heavy daily smokers (50 people). Traditional data filters will analyze all values within the attribute equally and may thus ignore the 5% of heavy smokers due to their relatively small size. However, when checking the instances with the heavy smoker value, we note that almost all people develop cancer (say 45 of 50 people) but the cancer rates within the remaining 95% (950 people) are not elevated. Thus, assume even a complete entropy reduction within that 5%, but no entropy reduction in the remainder, that attribute’s information gain will be quite small and selection algorithms using this measure as its criteria will fail to see its significance. Similarly, even selection measures not based on information gain will typically not note the significance of this subset of minority attribute values due to their size. We provide an overview of selection algorithms in the next section.

As we also describe in the next section, the medical importance of how these minority subsets may be correlated to a person’s genomic disposition leading to susceptibility or resilience for disease. Consequently these values can have strong medical diagnostic implications. The MIAT algorithm defined in this paper quantifies an approach of reasoning about whether the minority set of instance values is too small, and thus potentially overfitted, and just how strong an indication this group of values may be. Once important minority values are identified, we posit that we can potentially significantly improve our prediction model.

To demonstrate the effectiveness of MIAT, we present results from a gastrointestinal study that motivated this work. We found that MIAT was able to significantly improve the model’s predictive ability from 72.2% to 86.67%. More importantly, a cost biased model using MIAT was able to achieve a sensitivity of 0.92 while having a specificity of 0.6. This model identifies risk [17]. Additionally, to verify the generality of MIAT, we considered a canonical UCI Hungarian Heart Disease Dataset[3] to benchmark MIAT’s performance. Here again, we found that MIAT improved the model’s accuracy from 83.5% to 90.1% which also improving the sensitivity from 0.66 to 0.87.

This paper is organized as follows. In the next section, we present an overview of feature selection techniques to help stress the novelty of MIAT. We also present related genomic work which may provide a biological explanation for why this specific technique works. In Section III we present the methodology used in our gastrointestinal study and detail which attributes were collected. Section IV generally presents the MIAT algorithm and provides examples in the context of our study. Section V details how we applied both MIAT and other feature selection approaches and details its success over the current state-of-the-art in gastrointestinal cancer detection [17] as well as in the canonical UCI dataset. Last, we conclude and provide future research directions.

II. RELATED WORK

Attribute selection (also called feature selection) has been generally shown to help within medical domains [16]. In the context of medical research, the primary advantages to using attribute selection are that it: a) helps improve performance by overcoming the curse of dimensionality, b) provides faster modeling and can help medical practitioners avoid collecting unnecessary data, and c) assists in advancing the medical field by understanding which attributes are most important for an accurate diagnosis [15], [16].

Generally, these approaches can be divided into three categories: filters, wrappers and embedded approaches. Filters operate by creating a score value based on the attributes’ connection to the target values. A threshold is typically then implemented whereby attributes with a lower score are removed. Additionally, filter methods can be used to consider subsets of multiple features together and their correlation to the target variable [15]. Examples of single-variable filters include Euclidian Distances and Information Gain measure [16], which consider the entire set of all attributes’ values and their connection to the goal variable. In the context of medical diagnostic problems, Lovell et al. have addressed attribute selection in a problem of predicting adverse outcome in pregnancy using maternal features [18]. They show how a binary representation of features has substantial overlaps across classes, and that features may be ranked in terms of their likelihoods to define a maximum attainable discrimination, measured by the area under a receiver operating characteristic (ROC) curve. Even methods that consider multiple variables, such as correlation feature selection (CFS) [19], weigh all values within every attribute equally. Consequently, only the correlation between two or more attributes’ full range of values are considered in regard to those within the target class.

While filter methods are independent of the exact data mining algorithm under consideration, both wrappers and embedded methods link the selection process to the algorithm used to build the model. Within the wrapper approach, subsets of attributes are trained and tested while considering a specific classification model. The term “wrapper” indicates that the search for the optimal attribute set is “wrapped” around the classification model, thereby leaving the classification method with its classic implementation. In contrast, embedded methods link the selection process and model learning even further and perform the selection as part of the model training process [15], [16]. Both of these approaches have the disadvantage of being tightly interwoven to the algorithms being considered, making it difficult to generalize any general information, here medical diagnostic information, from the filtering stage. Furthermore, filtering has the advantage of being able to be performed only once, and then allowing all future models to be created and evaluated based on these findings.

To the best of our knowledge, the uniqueness of MIAT lies in its analysis of its approach of considering a score

---

that measures the values of even a subset of values for a given attribute. Thus, we consider each subset of values within a given attribute and compute these attributes’ general expressiveness independently of the other values within the attribute. As support for this approach, we posit that there is substantive empirical evidence that in other problems with data in high dimensions, the effective discriminant information resides in a small subspace, or subset of the variables. Li and Niranjan (2007) \cite{20}, for example, found patterns in such subsets in problems of microarray classification of gene expressions, text analysis and drug-like properties of synthetic molecules predicted from chemical fingerprint representation. They used a greedy sequential forward selection algorithm (wrapper) to show that beyond about ten percent of the features generalization ability saturates or, even worse, decreases. However, they did not consider applying such approaches to attribute selection as this paper does.

There seems to be emerging support for considering important subsets of attributes as being either “triggers” or “indicators” for biological processes. Recent genomic (DNA) and transcriptomic (RNA) research has shown that some people may have a natural predisposition or immunity towards certain diseases. Susceptibility at the genomic level is caused by mutations in various genomic loci, whereas at the transcriptomic level is caused by over or under activating cellular pathways. On the negative side, it seems that certain people have predispositions for certain diseases that can be triggered by certain events, for example the multistep mutations involved in the progression to colorectal cancer \cite{21} or the over activating of immune response and NF-kB pathways that causes chronic gastritis which is an inflammatory condition linked to the development of MALT lymphoma \cite{22}. Fortunately, these predispositions are not always realized and an open research question is what, if anything, triggers these dispositions. At the other extreme, it seems that certain people are inherently less susceptible to disease based on their genomic signature, and thus we would expect to see positive manifestations of this. A good example of this is the gain of function single nucleotide polymorphisms (SNPs). For example, the IL-23R R381Q mutation seems to protect against pediatric Crohn’s disease \cite{23} and mutations in TRPC4 cation channel gene protects against myocardial infarction \cite{24}. We posit that the success of MIAT is in finding the significant cases, albeit even ones that occur in a minority of instances, that trigger or indicate these changes. Thus, we shift our focus from studying all values for attributes, to finding those important subsets of values within a given attribute. Based on this genomic background, we would expect these subsets to be typically found at the extremes of behavior, that are potentially triggering transcriptomic changes or are indicative of an inherent genomic resilience.

III. System Implementation and Methodology

The general methodology behind this study was to gather a large set of patient data and to then create a predictive model for UGI cancer. The study was approved by the University’s local ethics committee. Participants provided written, informed consent and completed our study questionnaire. Initial work involved utilizing data from a previous study (BEST2 study) \cite{17} where 472 patients were recruited and provided data on their disease specific risk factors and symptoms. We validated the learned models on a second dataset of patients aged between 18-91 attending the University for an UGI endoscopic examination. This second dataset utilized a specifically designed questionnaire, outlined below, and is referred to as the UCL dataset. In total, 151 patients were recruited in this group with all patients providing in-depth risk factor and symptom data prior to undergoing an endoscopic examination of their UGI tract so their diagnosis could be definitively recorded.

Our primary aim was to predict which individuals are at risk of developing oesophageal or gastric cancer. As such, in the UCL dataset the patients were split into two groups; “at-risk” and “no-risk”. At risk lesions were identified through a literature search and includes all lesions that incur a risk of cancer progression to an individual. These include Barrett’s Oesophagus; oesophageal squamous dysplasia; gastric intestinal metaplasia and dysplasia; and chronic atrophic gastritis \cite{8}, \cite{25}, \cite{26}. Patients with these lesions or established cancers were included in the “at-risk” group. All others were placed in the “no risk” group. For this analysis 52 of the patients recruited were subsequently found to have normal UGI tracts (“no risk”) and 38 were identified to be “at-risk”.

People recruited as part of the UCL dataset filled out a specifically designed in-depth medical questionnaire. Questions included patient demographic information and physical measurements, such as waist / hip ratio. Other questions asked people to quantify if they could be identified as belonging to risk factors thought to be of importance in the development of UGI cancer as per previous publications \cite{17}. When designing our questionnaire we assimilated data collected from the BEST2 study as well as risk factors and symptoms linked to UGI cancer identified following a thorough literature search. The additional data collected explored risk factors such as dietary habits and those related to a patients quality of life. Our questionnaire explored risk factors at a greater depth than previous studies. For example, patients were not just asked what their average number of cigarettes smoked per day was but to identify their average cigarettes consumption at various ages (<16, 16-25, 26-35, 36-45, 46-60, >60 years old) thus providing a lifetime overview. Where possible, previously validated questions were used and the questionnaire was completed in the presence of a research nurse or clinical fellow in order to ensure accuracy. In total, we logged 205 unique attributes for each patient, which we then used as input to create the predictive model.

IV. The MIAT Algorithm

As one may assume, a patient database containing 205 attributes can and did help us create an accurate a predictive model. However, given the large number and range of value for each of these features, we assumed that we would encounter a “curse of dimensionality”. As such, and as per previous work on attribute selection, we assumed that “standard” feature selection approaches would help improve our predictive model \cite{15}, \cite{16}, \cite{19}, as Section \textbf{V} does in fact demonstrate. However, we noted that at times certain specific attributes’ values, and particularly those within the patient questionnaire, were significantly more valuable than others and contained strong indications if a person would develop cancer.

To quantify this phenomenon, we define the problem as follows: We assume that n attributes exist in the dataset, which
are denoted $A_1 \ldots A_n$. Each given attribute, $A_j$, has a set of $m$ discrete values, which we denote $val_1 \ldots val_m$. While each attribute, $A_j$, could potentially have different values for every database instance, making $m$ very large, $m$ was typically quite small for the attributes we considered (e.g. $m<=6$). Potentially, attributes with large values of $m$ could be redistributed and clustered around a smaller, fixed value for $m$. Similarly, attributes with continuous variables (e.g. the number of cigarettes they had smoked recently, or their weight) could be discretizing through existing techniques [27]. We also assume that the collection of all attributes provides an indication to one of the two values in the target variable, $T_1$ or $T_2$. Within medicine, this assumption typically holds as most learning tasks involve only two values in the target category (e.g. “at-risk” or “no risk”). However, more generally, we write the algorithm assuming that $x$ values in the target variable exist such that it checks for any strong indications by looping through all $x$ possible target values of $T_1 \ldots T_x$.

Based on these definitions, Algorithm 1 presents the MIAT attribute selection approach. In lines 1 - 3 of the algorithm, we loop through all attributes within the dataset and consider each subset of values for that given attribute, $A_k$, for all target variable values $T_1 \ldots T_x$. The algorithm will then check if the values within $val_k$ are significant and should be identified by our algorithm. To measure significance we require two conditions. First, we require that the size of the subset of values in $val_k$ not be overly small. This is done to prevent overfitting, and we generally require that this size be larger than a minimal threshold of $\epsilon$ (line 4). In our implementation, we assume this size was relative to the dataset size, not an absolute one. Second, this subset must have a strong indication to one of the values for the independent variable we are trying to learn. Towards this goal, we present a second threshold $\gamma$ (line 5). As we use entropy for evaluating if the subset is a strong indicator, we require small values of entropy as this indicates a strong indication towards the target’s value.

**Algorithm 1 MIAT Algorithm to Identify Attributes with Important Values**

1: for $i = T_1$ to $T_2$ do
2: for $j = A_1$ to $A_n$ do
3: for $k = val_1$ to $val_m$ do
4: if $Size(val_k) > \epsilon$ then
5: if $Entropy(val_k) < \gamma$ then
6: Identify $A_j$ as Important
7: end if
8: end if
9: end for
10: end for
11: end for

Several things are important to note about this algorithm. First, we assume that $m$ is greater than 2. If $m$ is 2, then “standard” feature selection approaches will work equally well as there will be no difference between the attributes MIAT selects versus those through other single variable attribute selection approaches, and especially those based on information gain scoring measures. Conversely, if $Size(val_k)$ is large, then MIAT is also unlikely to behave differently than existing approaches. This is because if a large subset has low entropy, then standard approaches should also flag $A_i$ as being worthy as $A_i$’s overall information gain is likely to be large due to the weight of $val_k$ within $A_i$. The true added value of MIAT is when $val_k$ is a small subset with low entropy, yet this subset is still not too small its selection overfitted. In these cases alone MIAT will still flag $A_i$ as being worthy of selection, while all previous feature selection algorithms we know of will not choose this attribute. Third, please note that line 5 considers the entropy of each subset of attribute values separately. As such, if all people within a certain subset, such as people who are extreme smokers, develop cancer (e.g. target value $T_1$), then its entropy is low (e.g. $Entropy(val_k) = 0$) and it will thus be selected. On the flipside, if all people within a certain subset, say people whom have had no heart burn within the past 12 months, have no cancer (e.g. target value $T_2$), than their entropy is again zero, as again all instances within $val_k$ belong to the same target variable value. Thus, MIAT will again select this attribute as being significant as $Entropy(val_k) < \gamma$. It is also important to note that while we use entropy here as the selection criteria, other measures such as lift, could potentially be used instead.

Figure 1 presents a small sample dataset and how which attributes MIAT will identify as being significant. Please note that attributes $A_1$ and $A_2$ do not yield low entropies in regard to values $T_1$ and $T_2$ and thus are not selected by the algorithm. Conversely, attribute $A_4$ is the trivial case which would be selected by MIAT, but as all values within $A_4$ yield low entropies, MIAT’s identification is the same as traditional feature selection through information gain (denoted through pink coloring). Within $A_3$, the subset values of $X$ and $Z$ both indicate a target value of $T_1$. However, as these values do not have a low entropy in $T_1$, they will not be selected by the algorithm. In contrast, while the subset of $Y$ values in $A_3$ also contains one instance in the $T_1$ target value, its entropy is only zero for $T_2$ (yellow boxes) and thus MIAT will identify $A_j$ as important based on this low value.

One motivation for MIAT comes from the observation that the probability distribution of features relating to any medical classification problem need not be uni-modal. In fact, as discussed in the previous smoker example, features are likely to be multi-modal with some modes having very low representation in the sampled dataset, but within those modes the prevalence of one class is significantly different from that for the remainder of the samples. It is known in standard pattern classification theory [28] that uni-modal data, particularly Gaussian data with equal covariances leads to simple classification boundaries while multi-modality leads to complex non-linear boundaries in the input space. This is precisely what sophisticated machine learning algorithms such as multi-layer neural networks and kernel support vector machines are designed to capture using their universal approx-

### Figure 1

**A Sample Dataset with MIAT Applied.**

<table>
<thead>
<tr>
<th>Person</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>New A3</th>
<th>A4</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person2</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Not-Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person3</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>Not-Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person4</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Not-Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person5</td>
<td>X</td>
<td>Z</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>T2</td>
</tr>
<tr>
<td>Person6</td>
<td>Z</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>T2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>New A3</th>
<th>A4</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person2</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Not-Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person3</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>Not-Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person4</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Not-Y</td>
<td>Y</td>
<td>T1</td>
</tr>
<tr>
<td>Person5</td>
<td>X</td>
<td>Z</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>T2</td>
</tr>
<tr>
<td>Person6</td>
<td>Z</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>T2</td>
</tr>
</tbody>
</table>
MIAT is designed to identify such low representation modes and separates them into additional features (see Algorithm 1). In our implementation, as shown in the figure, this will mean that we will add an additional column which is a boolean value of true if $val_k$ exists in $A_j$, and false otherwise for all subset values not $val_k$. This might seem strange as this is actually the opposite of feature reduction as we add new features where important attribute subsets are found. As we demonstrate in the results section, these added features may or may not be later removed by other feature selection approaches. However, their very existence does help overcome the curse of dimensionality by identifying these important portions of the attribute search space, thus facilitating more accurate models. As the under-represented modes of discriminant features are pulled out as new features, data that is separable along these new features facilitates improved accuracy. We now show empirical evidence in support of this, applied to two benchmark datasets from machine learning repositories and the gastrointestinal cancer classification problem of interest to us.

V. RESULTS

To study MIAT’s effectiveness, we analyzed three different types of studies within the two datasets under consideration. First, we analyzed which attributes were selected and why. Second, we created prediction models and noted that MIAT facilitated the creation of models that significantly outperformed those built without any attribute selection and even those with traditional single or multiple-variable filters \[19\]. Last, as our motivating problem was to effectively diagnose gastrointestinal cancers, we compared our results with the current state-of-the-art \[17\] and present how we built a model that outperforms in terms of both sensitivity and by nearly a three-fold factor of specificity.

Recall that one of the key advantages of attribute filtering over wrapper and embedded methods is in its ability to generally express what attributes are most worthy of consideration. To illustrate this point, we applied MIAT to the dataset we collected as per the methodology expressed in Section III. We applied MIAT with a minimum size of 0.10% of the dataset as per line 4 of Algorithm 1, and an entropy threshold of 0.2. This yielded 15 different values being added by the algorithm for a new total of 220 attributes. For the “at-risk” value of the target variable, we found that heavy smoking (from 4 options), heavy use of antacids (from 4 options), and heavy salt intake (from 5 options) were among the subset of values being selected by the algorithm. Strong indications for being in the “no-risk” value of the target variable included people who never had chest pains within the last 12 months (from 4 options), never took antacids (from 4 options) and never reported having their activity disrupted by stomach symptoms (from 5 options). While many of these risk factors, including antacid use and smoking are known risk factors, the medical experts we presented these findings to were intrigued by the existence of counter-categories (e.g. no antacid being related to no cancer). Similarly, the connection between activity and cancer was not previous known and will be studied further in the future.

We used the implementation of various data mining algorithms within the Weka data mining package \[29\] to assess the impact MIAT had on the overall accuracy, specificity and sensitivity in this dataset. All experiments were performed with 10-fold cross validation. While we considered several algorithms, as we wished to construct a probabilistic model of developing cancer, we constructed models based on Bayesian Networks unless noted otherwise.

The first 4 columns of Table I compare the overall accuracy, sensitivity, specificity and ROC for 6 different models built with the gastrointestinal cancer dataset. The first column represents the results with all 195 attributes that were collected. This column represents the baseline for all results without any attribute selection or the MIAT algorithm being applied. The second column represents the results with a total of 220 attributes—the 195 original ones with the 15 attributes MIAT added. Note that already here we see an improvement in results, despite the adding of this information. This demonstrates that MIAT has the potential to overcome the curse of dimensionality as evidently this data is not just “lost” in the added dimensionality of the search space. The third column presents the result with only the information selected by MIAT. Please note that while these results are significantly worse than those in the other datasets, this is not unexpected. As MIAT only selects attributes with small subsets of strongly indicative values, it is meant to augment, not replace, traditional methods. This is because our implementation of MIAT does not select the trivial cases where it would behave similar to more traditional single-variable filtering methods. We did this to demonstrate this very effect— that the MIAT approach of only selecting variables with small interesting subsets is alone not sufficient. In contrast, please compare the effect of attribute selection without the MIAT added attributes and those with (columns 4 and 5 respectively). While the CFS attribute selection approach did significantly improve the overall model accuracy, please note that this improvement was primarily in the model’s specificity (its accuracy to find the non-cancer patients). However, the more important sensitivity (recall of cancer patients) was the same. In contrast, MIAT in combination with CFS did only slightly better in its overall accuracy as its specificity was the same as that with CFS without MIAT, but the more important sensitivity result improved significantly. Similarly, please note that the ROC also improved nearly 10% to 0.89 showing the model’s improved ability to identify these cancer patients.

Previous work within a similar gastrointestinal dataset, but
with significantly fewer attributes, reported a sensitivity of 90% and specificity of just 22% [17]. While we believe the authors were correct in their desire to not miss significant amounts of cancer patients, we note that the 22% specificity indicates that the vast majority of those tested would still undergo an invasive examination. Please note that even with MIAT, our base model yielded a specificity of well above 90%, but with a sensitivity of only about 75%. However, given that the ROC of this model is quite high (see Figure 2), and much higher than the 0.64 that was previously reported [17], we reasoned that a cost sensitive model could be easily implemented that achieves the desired sensitivity while only slightly sacrificing the specificity. In order to implement such a model, we used the known Metacost algorithm [30] to create a cost bias away from false negatives. While we tested several algorithms within the Metacost algorithm, we found that the k-nn algorithm worked best in tandem with it (for k=3), the results of which are found in the last column of Table I. Please note that this combination yielded a sensitivity of 0.92 while achieving a specificity of 0.67 or over 3 times the previously reported result [17]. Thus, MIAT enabled us to achieve similar sensitivity to previous work while eliminating over 3 times the amounts of false positives.

In order to verify the generality of MIAT, also considered the canonical Hungarian Heart Disease Dataset from the UCI repository. We again applied MIAT and compared the overall accuracy, sensitivity, specificity and ROC for same 6 different models we built with the gastrointestinal cancer dataset. Once again, we used the Weka package [29], and here all models were constructed with Bayesian Networks.

Please note from Table II that again the combination of MIAT and CFS yielded the best results with MIAT primarily facilitating gains within the model’s sensitivity (recall of heart disease). The first column again represents the full dataset (13 attributes). The second column displays results after MIAT added 7 new attributes (20 total). The fourth and fifth columns again compare the results taken with CFS filtering on the original dataset, versus the results taken after filtering from with the MIAT attributes added. The last column again presents a cost-sensitive model, again using MetaCost [30] in conjunction with Bayesian Networks.

We believe several items are noteworthy from these results. First, we were somewhat surprised that CFS filtering actually slightly hurt the results over the initial database. We assume this may be due to the relatively small number of attributes in the dataset. Similarly, this may be why MIAT alone was successful. As MIAT presented 7 attributes, and the original dataset had only 14, it is possible that MIAT alone accurately covered the database’s attribute space without the original data. In contrast, the gastrointestinal cancer was much bigger and MIAT selected only 15 attributes of 205, which may explain why there MIAT alone was not sufficient. However, in both datasets we note that MIAT aided in finding the harder to diagnose minority cases of cancer and heart disease. Thus, we note that while MIAT does work with even smaller datasets, including the canonical set we considered here, we posit that it is best suited for large-dimension datasets where traditional filtering models fail to find important values within the minority value of the target variable.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we present MIAT, an attribute selection algorithm that finds important subsets of attributes’ values. As opposed to traditional attribute selection algorithms that consider all attribute values equally, MIAT explicitly considers each subset of values within an attribute separately and computes a score based on entropy to decide if this subset of values is significant. Motivated by genomic work [21]–[24], we posit that subset values that form the minority of instances, but are strong indicators for disease, can still be particularly important. As support for this thesis, we present results demonstrating that MIAT allowed us to significantly improve the accuracy of a gastrointestinal cancer dataset that motivated our work and performed significantly better that the state-of-the-art model [17]. To demonstrate the generality of MIAT, we also considered a canonical Hungarian Heart Disease Dataset where again MIAT facilitated improved results.

For future work, we are considering several important complimentary paths. First, we hope to incorporate genomic data within our study. By applying MIAT to this data we hope to uncover the genomic markers that make a person either particularly susceptible or resilient to forms of gastrointestinal cancer. Similarly, running accepted attribute selection techniques in conjunction with MIAT, and particularly those

<table>
<thead>
<tr>
<th></th>
<th>Full Data</th>
<th>Full Data with MIAT</th>
<th>MIAT Alone</th>
<th>Attribute Selection</th>
<th>Attribute Selection with MIAT</th>
<th>Cost Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>72.22%</td>
<td>75.56%</td>
<td>61.11%</td>
<td>83.33%</td>
<td>86.67%</td>
<td>77.78%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.66</td>
<td>0.68</td>
<td>0.27</td>
<td>0.68</td>
<td>0.76</td>
<td>0.92</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.77</td>
<td>0.79</td>
<td>0.87</td>
<td>0.94</td>
<td>0.94</td>
<td>0.67</td>
</tr>
<tr>
<td>ROC Area (AUC)</td>
<td>0.79</td>
<td>0.8</td>
<td>0.61</td>
<td>0.83</td>
<td>0.89</td>
<td>0.86</td>
</tr>
</tbody>
</table>

**TABLE I. RESULTS FROM THE GASTROINTESTINAL CANCER DATASET**

<table>
<thead>
<tr>
<th></th>
<th>Full Data</th>
<th>Full Data with MIAT</th>
<th>MIAT Alone</th>
<th>Attribute Selection</th>
<th>Attribute Selection with MIAT</th>
<th>Cost Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>83.50%</td>
<td>88.12%</td>
<td>86.47%</td>
<td>83.17%</td>
<td>90.10%</td>
<td>88.78%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.8</td>
<td>0.85</td>
<td>0.8</td>
<td>0.77</td>
<td>0.87</td>
<td>0.9</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.87</td>
<td>0.91</td>
<td>0.92</td>
<td>0.89</td>
<td>0.93</td>
<td>0.88</td>
</tr>
<tr>
<td>ROC Area (AUC)</td>
<td>0.91</td>
<td>0.94</td>
<td>0.91</td>
<td>0.9</td>
<td>0.95</td>
<td>0.95</td>
</tr>
</tbody>
</table>

**TABLE II. RESULTS FROM THE HEART DISEASE DATASET**

https://archive.ics.uci.edu/ml/datasets/Heart+Disease
related to correlation of attributes, should demonstrate when and where MIAT is in fact uncovering new relationships or confirming the influence of existing marker. Second we hope to study how MIAT can be made more scalable. As we currently loop through all target values, attributes, and subsets of attribute values, there potentially could be difficulty in scaling this algorithm to datasets with thousands or hundreds of thousands of attributes and instances. However, we believe that MIAT’s run-time could be improved by explicitly reasoning about the likelihood a given subset of attributes values will be significant. For example, we could prioritize looping through those attributes which have been previously identified as being significant, or at the extremes such as only people who are extremely heavy smokers, or those who never need stomach medication. Third, while we considered MIAT in the context of filtering algorithms, it may be equally applicable within embedded or wrapper approaches. We believe this may be particularly true when considering models that need to reason about cost sensitive classification, as can be the case in medical domains. Please note that our results utilized the Metacost algorithm which itself is a wrapper algorithm, albeit for classification and not attribute selection [30]. Thus, we hope to consider future integration of MIAT with wrapper selection algorithms. Last, and possibly most importantly, it should be noted that MIAT is a general algorithm that could potentially be applied to any medical dataset. Its potential impact is therefore in theory quite large, and further research is needed to study what modifications, if any, are needed to apply it to other datasets.

REFERENCES