Applications of Machine Learning in Cancer Prediction and Prognosis

Dr. S. Subbaiah [#1], Mr. S. Muruganandam [*2]
[#1]Associate Professor, Dept of CS,
Sri Krishna Arts and Science College, Coimbatore
[#2]Research Scholar, PG & Research Dept of CS& CA,
Vivekanandha College of Arts and Sciences for Women (Autonomous), Tiruchengode

***Abstract*:- Machine learning is a part of man-made brainpower that utilizes an assortment of measurable, probabilistic and enhancement methods that enables PCs to "learn" from past models and to recognize hard-to-perceive designs from enormous, loud or complex informational indexes. This capacity is especially appropriate to restorative applications, particularly those that rely upon complex proteomic and genomic estimations. Therefore, AI is much of the time utilized in malignant growth determination and discovery. All the more as of late AI has been applied to malignant growth guess and forecast. This last methodology is especially intriguing as it is a piece of a developing pattern towards customized, prescient medication. In amassing this audit we directed a wide study of the various kinds of AI strategies being utilized, the sorts of information being incorporated and the exhibition of these techniques in malignant growth expectation and anticipation. Various patterns are noted, remembering a developing reliance for protein biomarkers and microarray information, a solid predisposition towards applications in prostate and bosom malignant growth, and an overwhelming dependence on "more established" innovations such counterfeit neural systems (ANNs) rather than all the more as of late created or all the more effectively interpretable AI techniques. Various distributed investigations likewise seem to come up short on a proper degree of approval or testing. Among the better structured and approved examinations obviously AI strategies can be utilized to generously (15-25%) improve the precision of foreseeing malignant growth weakness, repeat and mortality. At an increasingly essential level, it is likewise obvious that AI is additionally improving our fundamental comprehension of malignant growth improvement and movement.**

Keywords : Cancer, machine learning, prognosis, risk, prediction

INTRODUCTION

Machine learning is not new to cancer research. Artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis for nearly 20 years (Simes 1985; Maclin et al. 1991; Ciccheti 1992). Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images (Petricoin and Liotta 2004; Bocchi et al. 2004) to the classification of malignancies from proteomic and genomic (microarray) assays (Zhou et al. 2004; Dettling 2004; Wang et al. 2005). According to the latest PubMed statistics, more than 1500 papers have been published on the subject of machine learning and cancer. However, the vast majority of these papers are concerned with using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In other words machine learning has been used primarily as an aid to cancer diagnosis and detection (McCarthy et al. 2004). It has only been relatively recently that cancer researchers have attempted to apply machine learning towards cancer prediction and prognosis. As a consequence the body of literature in the field of machine learning and cancer prediction/prognosis is relatively small (<120 papers).

AI isn't new to malignant growth examine. Counterfeit neural systems (ANNs) and choice trees (DTs) have been utilized in disease discovery and determination for almost 20 years (Simes 1985; Maclin et al. 1991; Ciccheti 1992). Today AI strategies are being utilized in a wide scope of uses running from recognizing and ordering tumors through X-beam and CRT pictures (Petricoin and Liotta 2004; Bocchi et al. 2004) to the grouping of malignancies from proteomic and genomic (microarray) measures (Zhou et al. 2004; Dettling 2004; Wang et al. 2005). As indicated by the most recent PubMed insights, in excess of 1500 papers have been distributed regarding the matter of AI and malignant growth. In any case, most by far of these papers are worried about utilizing AI strategies to recognize, group, identify, or recognize tumors and different malignancies. As such AI has been utilized fundamentally as a guide to malignancy conclusion and location (McCarthy et al. 2004). It has just been moderately as of late that malignancy specialists have endeavored to apply AI towards disease expectation and guess. As an outcome the collection of writing in the field of AI and disease forecast/guess is moderately little (<120 papers).

The key objectives of malignant growth forecast and guess are particular from the objectives of disease location and finding. In malignant growth expectation/anticipation one is worried about three prescient foci: 1) the forecast of disease weakness (for example chance evaluation); 2) the forecast of malignancy repeat and 3) the expectation of disease survivability. In the main case, one is attempting to anticipate the probability of building up a kind of malignant growth before the event of the sickness. In the second case one is attempting to foresee the probability of redeveloping malignancy after to the clear goals of the malady. In the third case one is attempting to anticipate a result (future, survivability, movement, tumor-tranquilize affectability) after the finding of the ailment. In the last two circumstances the accomplishment of the prognostic forecast is clearly needy, to some extent, on the achievement or nature of the conclusion. Anyway an infection guess can just come after a restorative conclusion and a prognostic expectation must consider something other than a straightforward finding (Hagerty et al. 2005).

For sure, a malignant growth guess commonly includes numerous doctors from various claims to fame utilizing various subsets of biomarkers and different clinical components, including the age and general wellbeing of the patient, the area and sort of disease, just as the evaluation and size of the tumor (Fielding et al. 1992; Cochran 1997; Burke et al. 2005). Normally histological (cell-based), clinical (quiet based) and segment (populace based) data should all be deliberately coordinated by the going to doctor to think of a sensible guess. In any event, for the most gifted clinician, this isn't anything but difficult to do. Comparative difficulties additionally exist for the two doctors and patients the same with regards to the issues of malignancy counteraction and disease defenselessness expectation. Family ancestry, age, diet, weight (stoutness), high-hazard propensities (smoking, overwhelming drinking), and presentation to natural cancer-causing agents (UV radiation, radon, asbestos, PCBs) all assume a job in foreseeing a person's hazard for creating disease (Leenhouts 1999; Bach et al. 2003; Gascon et al. 2004; Claus 2001; Domchek et al. 2003). Sadly these regular "large scale" clinical, ecological and conduct parameters by and large don't give enough data to make strong forecasts or visualizations. In a perfect world what is required is some quite certain atomic insights regarding either the tumor or the patient's very own hereditary make-up (Colozza et al. 2005).

With the fast improvement of genomic (DNA sequencing, microarrays), proteomic (protein chips, tissue clusters, immuno-histology) and imaging (fMRI, PET, miniaturized scale CT) advancements, this sort of atomic scale data about patients or tumors would now be able to be promptly gained. Atomic biomarkers, for example, physical transformations in specific qualities (p53, BRCA1, BRCA2), the appearance or articulation of certain tumor proteins (MUC1, HER2, PSA) or the substance condition of the tumor (anoxic, hypoxic) have been appeared to fill in as ground-breaking prognostic or prescient markers (Piccart et al. 2001; Duffy 2001; Baldus et al. 2004). All the more as of late, mixes or examples of numerous sub-atomic biomarkers have been seen as considerably more prescient than single part tests or readouts (Savage and Gascoyne 2004; Petricoin and Liotta 2004; Duffy 2005; Vendrell et al. 2005) If these sub-atomic examples are joined with large scale clinical information (tumor type, inherited angles, chance factors), the power and exactness of malignancy anticipations and expectations improves considerably more. Be that as it may, as the quantity of parameters we measure develops, so too does the test of attempting to understand this data.

Before, our reliance on full scale data (tumor, patient, populace, and natural information) by and large kept the quantities of factors sufficiently little with the goal that standard factual strategies or even a doctor's very own instinct could be utilized to anticipate disease dangers and results. Not withstanding, with the present high-throughput demonstrative and imaging advances we currently end up overpowered with handfuls or even many atomic, cell and clinical parameters. In these circumstances, human instinct and standard measurements don't by and large work. Rather we should progressively depend on non-customary, seriously computational methodologies, for example, AI. The utilization of PCs (and AI) in malady forecast and visualization is a piece of a developing pattern towards customized, prescient prescription (Weston and Hood 2004). This development towards prescient drug is significant, not just for patients (as far as way of life and personal satisfaction choices) yet additionally for doctors (in settling on treatment choices) just as wellbeing business analysts and approach organizers (in actualizing enormous scale disease counteraction or malignant growth treatment arrangements).

Given the developing significance of prescient prescription and the developing dependence on AI to make expectations, we trusted it would hold any importance with lead a nitty gritty survey of distributed examinations utilizing AI strategies in malignant growth forecast and guess. The plan is to distinguish key patterns concerning the sorts of AI strategies being utilized, the sorts of preparing information being incorporated, the sorts of endpoint forecasts being made, the kinds of tumors being contemplated and the general execution of these techniques in foreseeing malignant growth defenselessness or patient results. Strangely, when alluding to malignant growth forecast and guess we found that most examinations were worried about three "prescient" foci or clinical endpoints: 1) the expectation of disease powerlessness (for example chance evaluation); 2) the expectation of malignancy repeat and 3) the forecast of disease survivability.

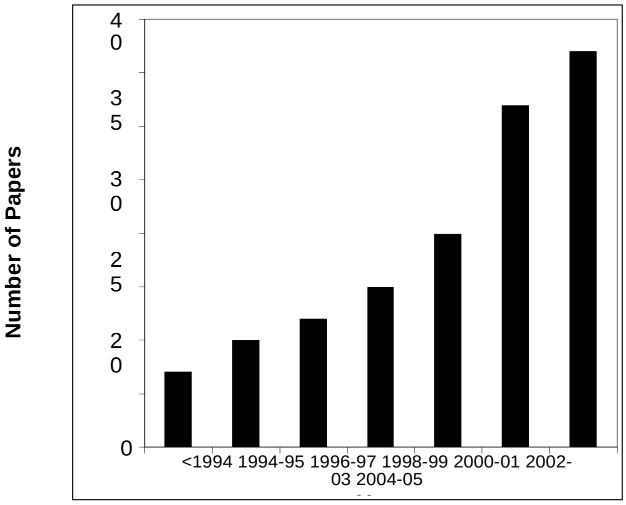


Figure 1. A histogram showing the steady increase in published papers using machine learning methods to predict cancer risk, recurrence and outcome. The data were collected using a variety of keyword searches through PubMed, CiteSeer, Google Scholar, Science Citation Index and other online resources. Each bar represents the cumulative total of papers published over a two year period. The earliest papers appeared in the early 1990's.

We likewise found that practically all expectations are made utilizing only four kinds of info information: genomic information (SNPs, changes, microarrays), proteomic information (explicit protein biomarkers, 2D gel information, mass otherworldly investigations), clinical information (histology, tumor organizing, tumor size, age, weight, chance conduct, and so on.) or mixes of these three. In looking at and assessing the current examinations various general patterns were noted and various normal issues recognized. A portion of the more evident patterns incorporate a quickly developing utilization of AI strategies in malignancy forecast and guess (Figure 1), a developing dependence on protein markers and microarray information, a pattern towards utilizing blended (proteomic + clinical) information, a solid inclination towards applications in prostate and bosom disease, and a startling reliance on more established innovations, for example, counterfeit neural systems (ANNs). Among the more normally noted issues was an irregularity of prescient occasions with parameters (too scarcely any occasions, an excessive number of parameters), overtraining, and an absence of outer approval or testing. By and by, among the better planned and better approved investigations obviously AI strategies, comparative with straightforward factual techniques, could generously (15-25%) improve the exactness of malignant growth vulnerability and disease result forecast. At the end of the day, AI has a significant task to carry out in malignant growth expectation and visualization.

Machine Learning Methods

Before starting with an itemized examination of what AI strategies work best for which sorts of circumstances, it is essential to have a decent comprehension of what AI is – and what it isn't. AI is a part of man-made brainpower explore that utilizes an assortment of factual, probabilistic and streamlining instruments to "learn" from past models and to then utilize that earlier preparing to arrange new information, distinguish new examples or foresee novel patterns (Mitchell 1997). AI, similar to insights, is utilized to break down and decipher information. In contrast to insights, however, AI techniques can utilize Boolean rationale (AND, OR, NOT), outright restriction (IF, THEN, ELSE), contingent probabilities (the likelihood of X given Y) and eccentric improvement methodologies to display information or arrange designs. These last techniques really look like the methodologies people ordinarily use to learn and group. AI despite everything draws intensely from measurements and likelihood, however it is essentially progressively incredible on the grounds that it permits derivations or choices to be made that couldn't generally be made utilizing regular factual approachs (Mitchell 1997; Duda et al. 2001). For example, numerous factual techniques depend on multivariate relapse or connection investigation. While by and large extremely amazing, these methodologies expect that the factors are autonomous and that information can be demonstrated utilizing direct mixes of these factors. At the point when the connections are non-direct and the factors are related (or restrictively reliant) traditional measurements normally wallows. It is in these circumstances where AI will in general sparkle. Numerous organic frameworks are in a general sense nonlinear and their parameters restrictively reliant. Numerous straightforward physical frameworks are direct and their parameters are basically autonomous.

Achievement in AI isn't constantly ensured. Likewise with any technique, a great comprehension of the issue and a valuation for the confinements of the information is significant. So too is a comprehension of the presumptions and confinements of the calculations being applied. In the event that an AI explore is appropriately planned, the students effectively executed and the outcomes heartily approved, at that point one for the most part has a decent possibility at progress. Clearly if the information is of low quality, the outcome will be of low quality (trash in = trash out). In like manner on the off chance that there are a larger number of factors than occasions to anticipate, at that point it is likewise conceivable to make a progression of repetitive students. This is a lot of learning calculations that appears to perform at the equivalent (low) level paying little heed to the decision of information. The issue of such a large number of factors and too hardly any models is known as the "scourge of dimensionality" (Bellman 1961). This revile isn't limited to AI. It influences numerous measurable strategies also. The main arrangement is to decrease the quantity of factors (highlights) or increment the quantity of preparing models. When in doubt, the example per-include proportion ought to consistently surpass 5:1 (Somorjai et al. 2003). Not exclusively is the size of the preparation set significant, so too is the assortment of the preparation set. Preparing models ought to be chosen to traverse an agent segment of the information the student hopes to experience. Preparing too often on too scarcely any models with too little assortment prompts the wonder of over-preparing or just preparing on commotion (Rodvold et al. 2001). An over-prepared student, much the same as an over-tired understudy, will for the most part perform ineffectively when it attempts to process or order novel information.

Now and again regular insights ends up being more dominant or more exact than AI. In these cases the client's underlying judgments about the relationship and nonlinearity of the information would have been off-base. This isn't really a shortcoming to AI, it is simply an issue of picking the correct apparatus for the correct activity. In like manner, not all AI strategies are made equivalent. Some are better for specific sorts of issues while others are better for different sorts of issues. For example some AI calculations scale pleasantly to the size of the natural areas, others don't. In like manner a few techniques may have suspicions or information necessities that render them inapplicable to the current issue. Knowing which technique is best for a given issue isn't intrinsically self-evident. This is the reason it is fundamentally imperative to attempt more than one AI strategy on some random preparing set. Another normal misjudging about AI is that the examples an AI apparatus finds or the patterns it recognizes are non-evident or not characteristically perceptible. Despite what might be expected, numerous examples or patterns could be identified by a human master – on the off chance that they looked hard enough at the information. AI basically saves money on the time and exertion expected to find the example or to build up the order conspire. Review that with any intriguing disclosure, it is as often as possible clear to the easygoing spectator – especially after the revelation has been made

There are three general sorts of AI calculations: 1) regulated learning; 2) solo gaining and 3) support learning. They are basically characterized based on wanted result of the calculation (Mitchell, 1997; Duda et al. 2001). In directed learning calculations a "farsighted supplier" or educator gives the learning calculation a named set of preparing information or models. These named models are the preparation set that the program attempts to find out about or to figure out how to outline input information to the ideal yield. For example a marked preparing set may be a lot of undermined pictures of the number "8" (Figure 2). Since every one of the pictures are named just like the number "8" and the ideal yield is the uncorrupted "8", the student can prepare under the supervision of an instructor mentioning to it what it should discover. This is the procedure by which most younger students learn. In unaided learning, a lot of models are given, however no names are given. Rather it is dependent upon the student to discover the example or find the gatherings. This is to some degree undifferentiated from the procedure by which most alumni understudies learn. Solo learning calculations incorporate such strategies as self-arranging highlight maps (SOMs), various leveled bunching and K-implies grouping calculations. These methodologies make bunches from crude, unlabeled or unclassified information. These bunches can be utilized later to create grouping plans or classifiers.

The SOM approach (Kohonen 1982) is a specific type of a neural system or ANN. It depends on utilizing a lattice of fake neurons whose loads are adjusted to coordinate information vectors in a preparation set. Truth be told, the SOM was initially intended to demonstrate natural cerebrum work (Kohonen 1982). A SOM starts with a lot of counterfeit neurons, each having its very own physical area on the yield map, which participate in a champ take-all procedure (a focused system) where a hub with its weight vector nearest to the vector of sources of info is proclaimed the victor and its loads are balanced making them closer to the information vector. Every hub has a lot of neighbors. At the point when this hub wins a challenge, the neighbors' loads are likewise changed, though to a lesser degree. The further the neighbor is from the champ, the littler its weight change. This procedure is then rehashed for each info vector for countless cycles. Various information sources produce various champs. The net outcome is a SOM which is fit for partner yield hubs with explicit gatherings or examples in the information informational collection.

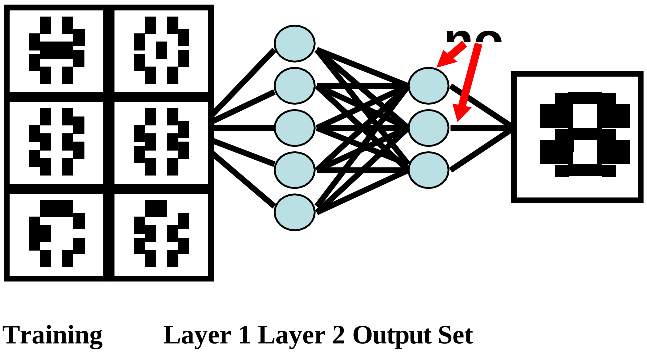


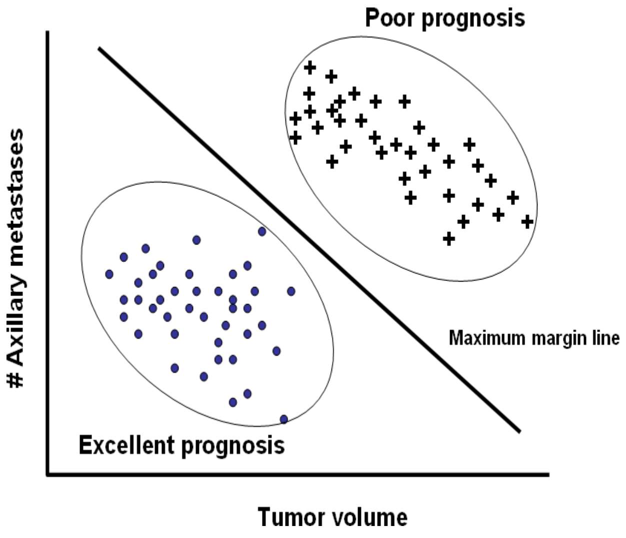
Figure 2. An example of how a machine learner is trained to recognize images using a training set (a corrupted image of the number "8") which is labeled or identified as the number "8".

The significant kinds of restrictive calculations include: 1) counterfeit neural systems (ANN – Rummelhart et al. 1986); 2) choice trees (DT – Quinlan, 1986); 3) hereditary calculations (GA – Holland 1975); 4) direct discriminant examination (LDA) techniques; 5) k-closest neighbor calculations anticipation with more than 820 of 1585 studied papers utilizing or alluding to ANNs. First created by McCulloch and Pitts (1943) and later promoted in the 1980's by Rumelhart et al. (1986), ANNs are fit for dealing with a wide scope of characterization or example acknowledgment issues. Their quality lies in having the option to play out a scope of factual (straight, calculated and nonlinear relapse) and legitimate tasks or inductions (AND, OR, XOR, NOT, IF-THEN) as a component of the arrangement procedure (Rodvold et al. 2001; Mitchell 1997). ANNs were initially intended to show the manner in which the mind works with various neurons being interconnected to one another through different axon intersections. Similarly likewise with organic learning, the quality of the neural associations is fortified or debilitated through continued preparing or fortification on marked preparing information. Scientifically, these neural associations can be spoken to as a wiring table or lattice (for example neuron 1 is associated with neuron 2, 4 and 7; neuron 2 is associated with neuron 1, 5, 6 and 8, and so forth.). This weight network is known as a layer, in relationship to the cortical layers in the cerebrum.

Neural systems commonly utilize various layers (called concealed layers) to process their information and produce a yield (Figure 2). To consent to the scientific structure of each layer, info and yield information is ordinarily organized as a string, or vector, of numbers. One of the difficulties in utilizing ANNs is mapping how this present reality input/yield (a picture, a physical trademark, a rundown of quality names, a guess) can be mapped to a numeric vector. In ANNs the change of neural association qualities is normally done by means of an improvement procedure got back to engendering (short for in reverse proliferation of mistakes – Rumelhart et al. 1986). This is a subsidiary based procedure that thinks about the yield of one layer to the previous layer's table. In straightforward terms the appropriate responses or marked preparing information are utilized to continuously adjust the numbers in the neural system's weight grids. Alearning or data move work (generally a sigmoidal bend) that is effectively differentiable is required for back proliferation. Most ANNs are organized utilizing a multi-layered feed-forward design, which means they have no input, or no associations that circle.

The plan and structure of an ANN must be altered or advanced for every application. Just picking a nonexclusive ANN engineering or gullibly organizing a conventional information/yield construction can prompt exceptionally horrible showing or very moderate preparing. Another hindrance of ANNs is the way that they are a "discovery" innovation. Attempting to make sense of why an ANN didn't work or how it plays out its characterization is practically difficult to perceive. As such, the rationale of a prepared ANN isn't anything but difficult to translate

As opposed to ANNs, the rationale of choice trees (DTs) is anything but difficult to observe. Officially a choice tree is an organized diagram or stream graph of choices (hubs) and their potential results (leaves or branches) used to make an arrangement to arrive at an objective (Quinlan, 1986; Mitchell 1997). Choice trees have been around for quite a long time (particularly in scientific classification) and are a typical segment to numerous medicinal symptomatic conventions. A framework of a straightforward choice tree for bosom malignant growth finding is given in Figure 3. Ordinarily choice trees are planned through conference with specialists and refined through long periods of experience or altered to consent to asset constraints or to confine hazard. Anyway choice tree students likewise exist which can consequently develop choice trees given a marked arrangement of preparing information. At the point when choice tree students are utilized to arrange information the leaves in the tree speak to groupings and branches speak to conjunctions of highlights that lead to those characterizations.



A choice tree can be scholarly by dynamically parting the marked preparing information into subsets dependent on a numerical or intelligent test (Quinlan 1986). This procedure is rehashed on each determined subset in a recursive way until further parting is either impractical, or a particular characterization is accomplished. Choice trees have numerous favorable circumstances: they are easy to comprehend and decipher, they require little information planning, they can deal with numerous kinds of information including numeric, ostensible (named) and clear cut information, they create powerful classifiers, they rush to "learn" and they can be approved utilizing factual tests. Anyway DTs don't for the most part proceed just as ANNs in progressively complex grouping issues (Atlas et al. 1990).

A to some degree more up to date AI procedure is known as a help vector machine or SVM (Vapnik, 1982; Cortes and Vapnik 1995; Duda et al. 2001). SVMs are outstanding in the realm of AI however practically obscure in the field of malignant growth expectation and guess (see Table 2). How a SVM functions can best be comprehended on the off chance that one is given a disperse plot of focuses, state of tumor mass versus number of axillary metastases (for bosom malignant growth) among patients with magnificent anticipations and poor visualizations (Figure 4). Two bunches are clearly apparent. What the SVM machine student would do is discover the condition for a line that would isolate the two groups maximally. In the event that one was plotting more factors (state volume,metastases and estrogen receptor content) the line of partition would turn into a plane.

On the off chance that more factors were incorporated the detachment would be characterized by a hyperplane. The hyperplane is controlled by a subset of the purposes of the two classes, called bolster vectors. Officially, the SVM calculation makes a hyperplane that isolates the information into two classes with the most extreme edge – implying that the separation between the hyperplane and the nearest models (the edge) is augmented. SVMs can be utilized to perform non-straight order utilizing what is known as a non-direct portion. A non-straight piece is a scientific capacity that changes the information from a direct component space to a non-straight element space. Applying various pieces to various informational indexes can drastically improve the presentation of a SVM classifier. Like ANNs, SVMs can be utilized in a wide scope of example acknowledgment and characterization issues extending from hand composing examination, discourse and content acknowledgment, protein work forecast and therapeutic determination (Duda et al. 2001). SVMs are especially appropriate to non-straight arrangement issues, as are k-closest neighbor draws near (see Table 1).

A Survey of Machine Learning Applications in Cancer Prediction

Figure 4. A simplified illustration of how an SVM might work in distinguishing between basketball players and weightlifters using height/weight support vectors. In this simple case the SVM has identified a hyperplane (actually a line) which maximizes the separation between the two clusters.

In setting up this survey a few electronic databases were gotten to including PubMed (biomedical writing), the Science Citation Index (biomedical , building, registering and physico-synthetic writing), CiteSeer (figuring writing), Google and Google Scholar (web-open logical writing). Question terms included "malignant growth and AI", "disease forecast and AI", "malignancy guess and AI", "malignant growth chance appraisal and AI" just as different sub-inquiries with explicit kinds of AI calculations. The significance of the individual papers was surveyed by perusing the titles and abstracts and distinguishing papers that utilized unmistakable AI strategies just as sub-atomic, clinical, histological, physiological or epidemiological information in doing a disease visualization or expectation. Papers that concentrated on determinations or basic tumor groupings were prohibited as were papers that had incidental appearances of the words "machine" or "learning" in their edited compositions. A PubMed search of "malignant growth and AI" yielded 1585 outcomes, while searches of "disease forecast and AI" and "disease guess and AI" yielded 174 and 240 hits separately. A point by point survey of these modified works prompted the distinguishing proof of 103 applicable papers of which 71 could be gotten to through different library possessions.

Utilizing CiteSeer, a hunt with the expressions "malignant growth and AI" yielded 349 outcomes, of which 12 (3.4%) were considered applicable to disease anticipation. Utilizing Google Scholar, an inquiry utilizing "malignant growth guess and 'AI'" yielded 996 outcomes, of which 49 (4.9%) were made a decision about pertinent to disease anticipation. A considerable lot of these papers were recently recognized in the PubMed look just like most by far of the hits in the Science Citation Index look. From the underlying gathering of papers recognized from these electronic inquiries, their reference records were additionally counseled to distinguish extra papers of intrigue or significance. At last in excess of 120 pertinent papers, going as far back as 1989, were recognized. Of these, 79 papers could be gotten to from existing library possessions and were chosen for increasingly nitty gritty examination (Table 2). While it is difficult to be sure that we accomplished total inclusion of all writing on AI and malignancy expectation/guess, we accept that a huge bit of the important writing has been evaluated for this audit.

Table 1. Summary of benefits, assumptions and limitations of different machine learning algorithms

Machine Learning Algorithm	Benefits	Assumptions and/or Limitations
Decision Tree (Quinlan 1986)	• easy to understand and efficient training algorithm • order of training instances	• classes must be mutually exclusive • final decision tree dependent upon order of attribute selection • errors in training set can result in overly complex

has no effect on training
decision trees
•
pruning can deal with the
□ missing values for an attribute make it problem of
overfitting
unclear about which branch to take when that
attribute is tested
Naïve Bayes
□ foundation based on
□ assumes attributes are statistically (Langley et al
statistical modelling
independent*
1992)
□ easy to understand
□ assumes normal distribution on numeric and efficient
training
attributes
algorithm
□ classes must be mutually exclusive
•
order of training instances
□ redundant attributes mislead classification has no effect
on training
□ attribute and class frequencies affect
•
useful across multiple
accuracy domains
earest □ fast classification of
□ slower to update concept description
Neighbour
instances
□ assumes that instances with similar attributes will (Patrick
&
□ useful for non-linear
have similar classifications
Fischer 1970;
classification problems
□ assumes that attributes will be equally relevant
Aha 1992)
□ robust with respect to
□ too computationally complex as number of attributes
irrelevant or novel
increases
attributes
•
tolerant of noisy instances or instances with missing attribute values
•
can be used for both regression and classification
Neural Network □ can be used for
□ difficult to understand
structure of algorithm (Rummelhart et al classification or
regression □ too many attributes can result in overfitting

1986)	• able to represent Boolean functions (AND, OR, NOT) • tolerant of noisy inputs • instances can be classified by more than one output	• optimal network structure can only be determined by experimentation
Support Vector Machine (Vapnik 1982; Russell and Norvig, p 749-52)	• models nonlinear class boundaries • overfitting is unlikely to occur • computational complexity reduced to quadratic optimization problem • easy to control complexity of decision rule and frequency of error	• training is slow compared to Bayes and Decision Trees • difficult to determine optimal parameters when training data is not linearly separable • difficult to understand structure of algorithm
Genetic Algorithm (Holland 1975)	• simple algorithm, easy to implement • can be used in feature classification and feature selection • primarily used in optimization • always finds a “good” solution (not always the best solution)	• computation or development of scoring function is non trivial • not the most efficient method to find some optima, tends to find local optima rather than global • complications involved in the representation of • training/output data

When taking a gander at the kinds of expectations or forecasts being made, by far most (86%) are related with foreseeing malignancy mortality (44%) and disease repeat (42%). In any case, a developing number of later investigations are presently planned for foreseeing the event of malignant growth or the hazard factors related with creating disease. When in doubt, paying little heed to the AI strategy utilized, the sort of expectation being made or the kind of malignant growth being assessed, AI techniques seem to improve the exactness of forecasts by and normal of 15-25% over other option or regular methodologies (Table 2)

In surveying how these forecasts were made apparently the larger part (53%) examines depended on clinical (malignant growth organizing, cell histology, atomic markers) or segment information (age, weight, smoking) – either alone or in mix with other sub-atomic biomarkers. While histological information is commonly increasingly open, the vagueness or pathologist-explicit idiosyncrasies of numerous histopathological appraisals quite often makes it hard to sum up or move an AI device prepared on this sort of information to other clinical settings. Given the constraints of utilizing histological evaluations in AI, there is an empowering pattern among later examinations to utilize all the more vigorously quantifiable highlights, for example, explicit protein markers, quality transformations and quality articulation esteems as information. Roughly 47% of concentrates utilized this atomic (for example proteomic or genomic) information either alone (25%) or in blend (22%) with clinical information. Given the accuracy of generally sub-atomic examines (with the conceivable exemption of microarray information), we accept the outcomes from these examinations ought to be all the more effectively or vigorously transferable to other clinical settings.

There is solid inclination among researchers to utilize AI towards anticipating results or dangers related with bosom (24%) and prostate (20%) malignant growth. This, almost certainly, mirrors the higher recurrence of these malignant growths among patients in Europe and North America. In any case, AI strategies seem to have been effectively utilized in anticipating results or dangers in almost twelve various types of malignancy. This proposes AI strategies can be commonly applied to malignant growth forecast and anticipation. Figure 5 likewise represents the circulation of the sorts of AI techniques applied to various types of disease forecasts. Practically 70% of every single revealed study utilize neural systems as their essential (and once in a while just) indicator. Bolster vector machines are an inaccessible second with 9%, while bunching and choice trees each record for about 6%. Hereditary calculations and different techniques (innocent Bayes, fluffy rationale) are seldom utilized (Table 2). This is both astounding and somewhat baffling. ANNs are generally old AI innovations which yield supposed "discovery" results. That is, their exhibition and arrangement forms are not effectively clarified or think. The presence of different strategies (SVMs, DTs, NBs) which innately give effectively open clarifications shows up not to be broadly known among malignancy informaticians. By and large, huge numbers of the papers audited for this study were of for the most part high caliber. A portion of the better papers are talked about in more detail under the "Contextual analyses" segment of this audit. Be that as it may, an upsetting number of concentrates needed adequate inner or outer approval, were prepared on excessively not many models, tried on just a solitary machine student or had no well-characterized standard with which to look at the exhibition of the announced calculation. These issues are examined in more detail under the area entitled "Restrictions and Lessons".

Lessons, Limitations and Recommendations

The 3 contextual investigations sketched out in the previous pages are only a couple of instances of how well-planned AI examinations ought to be led and how the strategies and results ought to be portrayed, approved and surveyed – particularly in malignant growth forecast and anticipation. There are clearly numerous different instances of similarly great investigations with similarly amazing outcomes (see Table 2). In any case, it is likewise essential to take note of that not all AI contemplates are led with a similar thoroughness or tender loving care likewise with these contextual investigations. Having the option to recognize potential issues in either the test plan, approval or student usage is basic not just for those wishing to utilize AI, yet in addition for those expecting to assess various examinations or to survey distinctive AI alternatives.

One of the most widely recognized issues seen among the investigations overviewed in this survey was the absence of consideration paid to information size and student approval. As it were, there are various examinations with messy test plan. A base necessity for any AI practice is having an adequately huge informational index that can be parceled into disjoint preparing and test sets or exposed to some sensible type of n-overlap cross-approval for littler informational collections. Normally 5-overlay (iteratively taking 20% of the preparation information out to fill in as testing information) or 10-overlap cross-approval (iteratively taking 10% of the preparation information out to fill in as testing information) is adequate to approve most any learning calculation. This sort of thorough inside approval is basic to making a vigorous student that can reliably deal with novel information. Past the standard act of inside approval, it is especially gainful to play out an approval test utilizing an outer information source. Outside approval is a significant "mental soundness" check and it likewise assists with getting or limit any inclination that might be forced by site or individual explicit clinical estimation rehearses. Obviously, this outer approval set should likewise be of adequately enormous size to guarantee reproducibility.

As has been much of the time noted previously, the size of a given preparing set has a few ramifications relating to heartiness, reproducibility and precision. The principal suggestion is that for a littler example size, practically any model is inclined to overtraining. Overtraining can prompt revealed exactnesses that might be deluding or wrong. For example, one early investigation revealed just a solitary misclassification during the preparation and testing of an ANN for anticipating the endurance of hepatectomized patients utilizing 9 separate highlights (Hamamoto et al. 1995). Be that as it may, the whole informational collection (preparing and testing) comprised of only 58 patients. This specific examination at that point utilized an outer informational index to approve the model where the creators tentatively anticipated the endurance result with 100% exactness.

Be that as it may, the outer test set just comprised of 11 patients. The way that 100% precision is achieved for a forthcoming forecast is amazing, yet given the size of the approval set and the little example per-highlight proportion, some uncertainty might be thrown on the heartiness of the indicator. Surely a bigger approval set would be alluring to fortify the case of 100% exactness. In another model, just 28 cases were utilized to assemble an ANN for foreseeing throat malignant growth repeat that utilized the articulation levels of 60 qualities from microarray information (Kan et al. 2004). The exactness of the model was professed to be 86%, yet this is especially speculate given the little example size. In reality all things considered, this ANN was over-prepared. The size of a given informational index additionally altogether influences the example per-include proportion. When in doubt, the example per-highlight proportion ought to be in any event 5-10 (Somorjai et al. 2003). Little example per-highlight proportions are an especially enormous issue for microarray considers, which frequently have a great many qualities (ie includes), however just several examples. The investigation by Ohira et al. (2005) gives one such case of the issues one may experience attempting to process an excess of microarray information. These creators made a probabilistic yield factual classifier to foresee guess of neuroblastoma patients utilizing microarray information from 136 tumor tests. Each microarray had 5340 qualities, prompting an example for each element proportion of ~0.025. An example for every component proportion this little is profoundly helpless to the issues of overtraining. Besides, with an example for every component proportion of this size it is likewise conceivable to grow exceptionally repetitive arrangement models which perform similarly well in spite of being prepared on various subsets of qualities. The issue with excess models is that the power of any one model can't be ensured as more experiments become accessible.

Data size is not the only limitation for effective machine learning. Data set quality and careful feature selection are also equally important (recall: “garbage in=garbage out”). For large data sets data entry and data verification are of paramount importance. Often careless data entry can lead to simple off-by-one errors in which all the values for a particular variable are shifted up or down by one row in a table. This is why independent verification by a second data-entry curator or data checker is always beneficial. Further verification or spot checking of data integrity by a knowledgeable expert, not just a data entry clerk, is also a valuable exercise. Unfortunately, the methods employed to ensure data quality and integrity are rarely discussed in most machine learning papers.

Just as data quality is important so too is feature quality. Certainly the subset of features chosen to train a model could mean the difference between a robust, accurate model and one that is flawed and inaccurate. Ideally features should be chosen that are reproducible and precisely measurable from one lab (or clinic) to the next. One study (Delen et al. 2005) used “primary site code” and “site specific surgery code” as features to predict breast cancer survivability. While these clinical features may be helpful in determining the outcome for breast cancer patients at this particular hospital, for this moment in time, they may become irrelevant over time. Even worse, if new site codes or site specific surgery codes are created, the model will have to be re-trained to account for the new codes. Similar feature selection problems often occur with histological assessments. As good as many pathologists are there is always some inconsistency (up to 30% in many cases) between different histopathological assessments from different sites or different pathologists. As a rule, the best features are those that are highly reproducible, universal or absolute (age, gender, weight, certain biomarker measurements, etc). Even with these seemingly robust features it is important to remember that clinical data sets are not static entities. With time the importance or relevance of these clinical measures may evolve over time with some features being added, modified or deleted. Therefore a classifier must also be able to adapt to different feature sets over time too.

Another important lesson that was learned from assessing many of these machine learning papers was the value of using multiple predictor models based on different machine learning techniques. While ANNs are often considered to be very sophisticated and advanced machine learning methods, ANNs are not always the best tools for the job. Sometimes simpler machine learning methods, like the naïve Bayes and decision tree methods can substantially outperform ANNs (Delen et al. 2005). Assessing the performance of a machine learning predictor against other predictors is critical to choosing the optimal tool. It is also critical to deciding if the method is any better than previously existing schemes. Ideally, any newly published machine learning model should be compared against either another kind of learning model, a traditional statistical model or an expert-based prognostic scheme such as the TNM staging system. As seen in Table 2, sometimes the more sophisticated machine learning methods do not lead to the best predictors. In some cases, traditional statistics actually outperform machine

It is likewise critical to recollect that the AI procedure is basically a computational test. Like any examination it depends on a theory, it follows characterized techniques and it expects information to be approved. Since machine students speak to genuine test systems, they ought to be treated in that capacity. Subsequently point by point methodological documentation is of vital significance. In a perfect world, the informational indexes utilized for preparing and testing ought to be portrayed in detail and madeaccessible to people in general. Data about preparing and testing information ought to likewise be well-depicted remembering the route for which the sets were parceled. Similarly the insights about the calculations utilized and their usage ought to be given or recorded to allow others to check and imitate the outcomes. On a fundamental level, the outcomes from a decent AI examination ought to be as reproducible as some other standard lab convention

Conclusion

In this audit we have endeavored to clarify, think about and evaluate the presentation of various AI that are being applied to disease expectation and forecast. Explicitly we distinguished various patterns as for the sorts of AI strategies being utilized, the sorts of preparing information being coordinated, the sorts of endpoint forecasts being made, the kinds of diseases being considered and the general execution of these techniques in foreseeing malignant growth defenselessness or results. While ANNs still prevail it is clear that a developing assortment of interchange AI methodologies are being utilized and that they are being applied to numerous sorts of tumors to foresee in any event three various types of results. It is additionally evident that AI techniques by and large improve the exhibition or prescient exactness of most visualizations, particularly when contrasted with regular factual or master based frameworks. While most examinations are commonly very much built and sensibly all around approved, surely more noteworthy consideration regarding test structure and execution seems, by all accounts, to be justified, particularly as for the amount and nature of organic information. Upgrades in exploratory structure alongside improved natural approval would no uncertainty improve the general quality, all inclusive statement and reproducibility of many machine-based classifiers. Generally, we accept that if the nature of studies keeps on improving, almost certainly, the utilization of AI classifier will turn out to be considerably more ordinary in numerous clinical and medical clinic settings

REFERENCES

[1] Aha D. 1992. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man- Machine Studies*, 36:267-287.

[2] Ando T, Suguro M, Hanai T, et al. 2002. Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma. *Jpn J Cancer Res*, 93:1207- 12.

[3] Ando T, Suguro M, Kobayashi T, et al. 2003. Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling. *Cancer Sci*, 94:906-13.

[4] Atlas L, Cole R, Connor J, et al. 1990. Performance comparisons between backpropagation networks and classification trees on three real-world applications. *Advances in Neural Inf. Process. Systems*, 2:622-629.

[5] Bach PB, Kattan MW, Thornquist MD, et al. 2003. Variations in lung cancer risk among smokers. *J Natl Cancer Inst*, 95:470-8. BaldusSE, Engelmann K, Hanisch FG. 2004. MUC1 and the MUCs: a family of human mucins with impact in cancer biology. *Crit Rev Clin Lab Sci*, 41:189-231.

[6] Bellman R. 1961. Adaptive Control Processes: A Guided Tour, Princeton University Press.

[7] Bocchi L, Coppini G, Nori J, Valli G. 2004. Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. *Med Eng Phys*, 26:303-12.

[8] Bollschweiler EH, Monig SP, Hensler K, et al. 2004. Artificial neural network for prediction of lymph node metastases in gastric cancer: a phase II diagnostic study. *Ann Surg Oncol*, 11:506-11.

[9] *Bottaci L, Drew PJ, Hartley JE, et al. 1997. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet*, 350:469-72.

[10] Bryce TJ, Dewhirst MW, Floyd CE Jr, et al. 1998. Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. *Int J Radiat Oncol Biol Phys*, 41:239-45.

[11] Burke HB, Bostwick DG, Meiers I, et al. 2005. Prostate cancer outcome: epidemiology and biostatistics. *Anal Quant Cytol Histol*, 27:211-7.

[12] *Burke HB, Goodman PH, Rosen DB, et al. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79:857-62.

[13] Catto JW, Linkens DA, Abbod MF, et al. 2003. Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro- fuzzy modeling and artificial neural networks. *Clin Cancer Res*, 9:4172-7.

[14] *Cicchetti DV. 1992. Neural networks and diagnosis in the clinical laboratory: state of the art. *Clin Chem*, 38:9-10.

[15] Claus EB. 2001. Risk models used to counsel women for breast and ovarian cancer: a guide for clinicians. *Fam Cancer*, 1:197-206.

[16] Cochran AJ. 1997. Prediction of outcome for patients with cutaneous melanoma. *Pigment Cell Res*, 10:162-7.

[17] *Colozza M, Cardoso F, Sotiriou C, et al. 2005. Bringing molecular prognosis and prediction to the clinic. *Clin Breast Cancer*, 6:61- 76.

[18] Cortes C, Vapnik V. 1995. Support-vector networks. *Machine Learning*, 20:273-297.

[19] Crawford ED, Batuello JT, Snow P, et al. 2000. The use of artificial intelligence technology to predict lymph node spread in men with clinically localized prostate carcinoma. *Cancer*, 88:2105- 9.

[20] Dai H, van't Veer L, Lamb J, et al. 2005. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res*, 65:4059-66.

[21] **De Laurentiis M, De Placido S, Bianco AR, et al. 1999. A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. *Clin Cancer Res*, 5:4133-9.

[22] *Delen D, Walker G, Kadam A. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med*, 34:113-27.

[23] Dettling M. 2004. BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20:3583-93.

[24] Domchek SM, Eisen A, Calzone K, et al. 2003. Application of breast cancer risk prediction models in clinical practice. *J Clin Oncol*, 21:593-601.

[25] Drago GP, Setti E, Licitra L, et al. 2002. Forecasting the performance status of head and neck cancer patient treatment by an interval arithmetic pruned perceptron. *IEEE Trans Biomed Eng*, 49:782- 7.

[26] Duda RO, Hart PE, Stork DG. (2001) Pattern classification (2nd edition). New York: Wiley.

[27] Duffy MJ. 2001. Biochemical markers in breast cancer: which ones are clinically useful? *Clin Biochem*, 34:347-52.

[28] *Duffy MJ. 2005. Predictive markers in breast and other cancers: a review. *Clin Chem*, 51:494-503.

[29] Dumitrescu RG, Cotarla I. 2005.Understanding breast cancer risk — where do we stand in 2005? *J Cell Mol Med*, 9:208-21.

[30] Ehlers JP, Harbour JW. 2005. NBS1 expression as a prognostic marker in uveal melanoma. *Clin Cancer Res*, 11:1849-53.

[31] Fielding LP, Fenoglio-Preiser CM, Freedman LS. 1992. The future of prognostic factors in outcome prediction for patients with cancer. *Cancer*, 70:2367-77.

[32] Fujikawa K, Matsui Y, Kobayashi T, et al. 2003. Predicting disease outcome of non-invasive transitional cell carcinoma of the urinary bladder using an artificial neural network model: results of patient follow-up for 15 years or longer. *Int J Urol*, 10:149- 52.

[33] **Futschik ME, Sullivan M, Reeve A, et al. 2003. Prediction of clinical behaviour and treatment for cancers. *Appl Bioinformatics*, 2(3 Suppl):S53-8.

[34] Gascon F, Valle M, Martos R, et al. 2004. Childhood obesity and hormonal abnormalities associated with cancer risk. *Eur J Cancer Prev*, 13:193-7.

[35] Grumett S, Snow P, Kerr D. 2003. Neural networks in the prediction of survival in patients with colorectal cancer. *Clin Colorectal Cancer*, 2:239-44.

[36] Gulliford SL, Webb S, Rowbottom CG, et al. 2004. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. *Radiother Oncol*, 71:3-12.

[37] **Hagerty RG, Butow PN, Ellis PM, et al. 2005. Communicating prognosis in cancer care: a systematic review of the literature. *Ann Oncol*, 16:1005-53.

[38] *Hamamoto I, Okada S, Hashimoto T, et al. 1995. Prediction of the early prognosis of the hepatectomized patient with hepatocellular carcinoma with a neural network. *Comput Biol Med*, 25:49-59.

[39] Hamilton PW, Bartels PH, Anderson N, et al. 1999. Case-based prediction of survival in colorectal cancer patients. *Anal Quant Cytol Histol*, 21:283-91.

[40] Han M, Snow PB, Epstein JI, et al. 2000. A neural network predicts progression for men with gleason score 3+4 versus 4+3 tumors after radical prostatectomy. *Urology*, 56:994-9.

[41] Hanai T, Yatabe Y, Nakayama Y, et al. 2003. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. *Cancer Sci*, 94:473-7.