Artificial Neural Networks: Predicting Head CT Findings in Elderly Patients Presenting With Minor Head Injury After a Fall

Michael W. Dusenberry, B.S.[a,1,*], Charles K. Brown, M.D.[b], Kori L. Brewer, Ph.D.[b]

[a] *Brody School of Medicine, East Carolina University*
600 Moye Blvd, Greenville, NC 27834, USA
[b] *Department of Emergency Medicine*
Brody School of Medicine, East Carolina University
600 Moye Blvd, Greenville, NC 27834, USA

Abstract

Objectives: To construct an artificial neural network (ANN) model that can predict the presence of acute CT findings with *both* high sensitivity and high specificity when applied to the population of patients $\geq$ age 65 years who have incurred minor head injury after a fall.
Methods: An ANN was created in the Python programming language using a population of 514 patients $\geq$ age 65 years presenting to the ED with minor head injury after a fall. The patient dataset was divided into three parts: 60% for "training", 20% for "cross validation", and 20% for "testing". Sensitivity, specificity, positive and negative predictive values, and accuracy were determined by comparing the model's predictions to the actual correct answers for each patient.
Results: On the "cross validation" data, the model attained a sensitivity ("recall") of 100.00%, specificity of 78.95%, PPV ("precision") of 78.95%, NPV of 100.00%, and accuracy of 88.24% in detecting the presence of positive head CTs. On the "test" data, the model attained a sensitivity of 97.78%, specificity of 89.47%, PPV of 88.00%, NPV of 98.08%, and accuracy of 93.14% in detecting the presence of positive head CTs.
Conclusions: ANNs show great potential for predicting CT findings in the population of patients $\geq$ 65 years of age presenting with minor head injury after a fall. As a good first step, the ANN showed comparable sensitivity, predictive values, and accuracy, with a much higher specificity than the existing decision rules in clinical usage for predicting head CTs with acute intracranial findings.

Keywords: Neural Network Models; Elderly; Head Injury, Minor; Falls

1. Introduction

Current evidence suggests that patients $\geq$ 65 years old presenting to the emergency department (ED) with minor head injury after a fall should receive a head CT scan. However, only a small percentage of these patients are actually found to have acute findings associated with the scan. Therefore, predictors for this class of patients that are *both* sensitive and specific would be desirable.

In 2004, injuries resulted in 31 million ED visits, representing 32% of all visits to the ED for any reason [1, 2]. Elderly patients are at the highest risk for both fatal and nonfatal injuries, with mortality and hospitalization rates for injuries reported to increase dramatically [1, 2]. Falls are the most common mechanism of injury for older patients visiting the ED, and are the most common cause of injury-related death [1, 2]. Due to the generally increased incidence of injury, specifically closed head injury, head CT is frequently ordered [3]. However, CT scans are costly and are recognized to carry a radiation risk [4, 5]; specifically, head CTs obtained because of a fall account for the expenditure of millions of dollars annually in the United States [6].

In 2001, a study was published to determine predictors of positive CT findings for patients of all ages with minor head injury, resulting in a highly sensitive decision rule known as the Canadian CT Head Rule (CCHR) [6]. Notably, an age $\geq$ 65 was a sensitive predictor of positive CT findings, but this age group was not further stratified.

Several other widely noted evidence-based decision rules for the general population [6, 7, 8] also indicate that age above 60 or 65 years places the patient at high risk for an abnormal head CT after mild head injury. These various decision rules have been compared to determine if one or

*Corresponding author
Email addresses: dusenberrymw@gmail.com (Michael W. Dusenberry, B.S.), browncha@ecu.edu (Charles K. Brown, M.D.), brewerk@ecu.edu (Kori L. Brewer, Ph.D.)
[1] Present Address: San Bruno, CA 94066, USA
Abbreviations - ANN: Artificial Neural Network

another more readily identifies the patient who will benefit from head CT [9, 10, 11, 12, 13, 14, 15], but none specifically addresses the population of patients over age 65 who potentially have an intracranial injury, particularly after a fall or other relatively minor mechanism. Currently, no definitive evidence exists regarding how to evaluate elderly patients after a fall, although a promising paper recently has been published using the non-age related New Orleans Criteria [16], predicting 100% of the abnormal head CT scans in an elderly population in a retrospective fashion.

A recent retrospective study [17] of 2149 elderly patients older than 65 years presenting to the ED with minor head injury found that 2.18% (47) of these patients had pathological findings, with only 0.14% (3) requiring neurosurgical intervention. Thus, while age ≥ 65 is inherently a 100% sensitive predictor for patients in the ≥ 65 age group, it has a low positive predictive value (PPV) due to only a small percentage of these positively-predicted patients actually having positive findings. Additionally, we note for clarity that the specificity is inherently 0% due to predicting all patients in this age group to be positive, thus missing all negative cases.

To contain costs while providing excellent care, it is important for emergency physicians to know if and when a patient will benefit from usage of a head CT scan.

1.1. Artificial Neural Networks

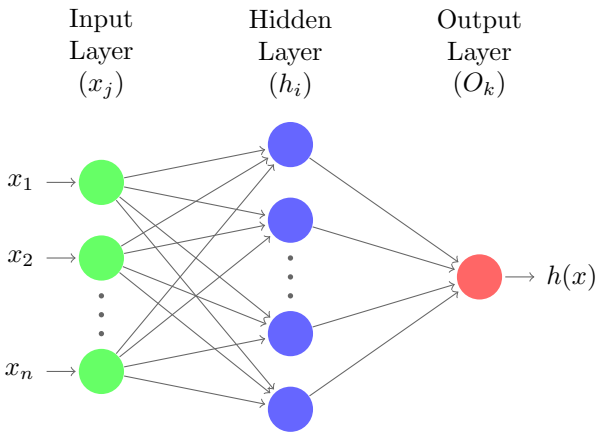

Figure 1: Artificial Neural Network Diagram

Artificial neural networks (ANNs) are mathematical models that are capable of learning highly complex, nonlinear relationships between a given set of input features [18, 19]. Due to their complex learning abilities, ANNs are widely used and studied in both practical and theoretical computer science research in machine learning and artificial intelligence. Given the ability to learn complex, nonlinear patterns from data, ANNs have great potential in being able to predict outcomes in complex medical cases.

As an example, a study was performed in 2001 to determine the effectiveness of using ANNs to predict CT findings in pediatric patients presenting with head trauma [20]. The authors collected data for pediatric patients who had presented to the ED with closed head injury and subsequently received a head CT scan, and then used this data to compare the performance of an ANN, a logistic regression model, and clinical judgment by pediatric EM physicians in predicting CT findings. Notably, the ANN was found to be both significantly more sensitive and more accurate than clinical judgment alone, with almost equal specificity. The ANN was also more sensitive than the logistic regression model.

Appendix A discusses ANNs in more depth, with specific regard to the type and formulation used in this study, and uses information gathered from various sources [18, 21, 22, 23]. The interested reader is encouraged to consult this appendix and the related sources for further information.

1.2. Objectives

The primary objective was to build a preliminary ANN model that could predict, with *both* high sensitivity and high specificity, the presence of CT findings in patients ≥ 65 years old who presented to the ED with minor head injury after a fall.

2. Methods

2.1. Study Design

A retrospective cohort study was performed. All patient records collected were de-identified prior to continuing with the study. The study was reviewed as UMCIRB #08-0773 and deemed exempt by the University and Medical Center Office for Human Research Integrity. Patient privacy and confidentiality of medical record information were the only ethical considerations deemed necessary.

2.2. Study Setting & Population

A retrospective chart review obtained 514 patients ≥ 65 years old presenting or transferring to the ED of a Level 1 Trauma Center teaching hospital with minor head injury after a fall between January 1st 2008 and September 30th 2013. The hospital is located in the Southeastern United States, with an ED volume of 90,000 patients during 2008, rising to approximately 120,000 in 2013 by the end of the study. To be eligible for inclusion, patients must have presented or transferred to the ED or its Fast Track area during the indicated time frame with minor head injury after a fall of any sort, including major trauma, and subsequently received a head CT that was interpreted for presence of acute findings. The final diagnosis must have contained ICD-9 codes for "fall" or "traumatic injury" (958.0-959.0). Physician judgment and standard accepted medical practice determined whether a patient received a head CT scan.

Prior to collection of study data, ten charts were randomly selected and all investigators extracted the prescribed data from each chart. Comparisons of the data

obtained by each investigator were made to assess consistency in interpretation of patient records and findings. The kappa statistic for inter-rater reliability was 0.86 and demonstrated good reliability. Collection of 514 patients required review of 5 years of medical records to obtain 227 positive findings. Since the prevalence rate of positive cases for this patient population is quite low, choice-based sampling [24] was used in which approximately equal numbers of patients for each class (positive vs. negative acute findings) were purposefully gathered in order to avoid skewed classes. Rare classes can lead to poor learning in ANNs and other related algorithms [25], and we discuss this further in section Appendix A.8. Any patients with missing data points were discarded, resulting in the final number of 514 patients.

2.3. Measurements

For each patient, several pieces of data were collected to be used as "features", including: gender; presence of dementia; use of aspirin or anticoagulants; presence of injury above the clavicle; type of fall; and presence of acute findings on head CT. Dementia was noted from the patient's past medical history, or from the current provider's note. A patient was considered to have a memory deficit if he or she had a change from their baseline memory status. Anticoagulants were categorized as: aspirin, clopidogrel, warfarin, fractional based or low molecular weight heparin. Presence, location, and type of injury were noted from the physician's note, and the discharge or admission diagnoses recorded in the chart for that visit. Trauma above the clavicles was considered as any physical evidence of trauma above the clavicles. The type of fall was characterized as from a bed, from sitting, from standing, or from a greater height. If the treating physician was unable to obtain any information, it was noted as "unable to obtain", and subsequently, these patients were removed. The official radiologist readings were used to assess presence of abnormal head CT. The word "acute" needed to appear in the radiology report describing the intracranial findings in order for the image to be considered "positive". The neurosurgical intervention rate was not recorded for this set of patients. Each of the features was numerically coded as either present or absent.

2.4. Data Analysis & Study Protocol

An ANN of the type thoroughly described in Appendix A was created and trained in the Python programming language using the Keras [26] neural network library, and code was written to process the patient datasets, build and train the ANN models, identify ideal training settings ("hyper-parameters"), and calculate relevant statistics, as reported in section 3.

In this study, the presence of acute CT findings was treated as the target outcome ("class"), and all other features were treated as inputs. The dataset was randomly divided into three parts for use in "training" (60%), "cross validation" (20%), and "testing" (20%). Models were built by learning on the training set and validating on the cross validation set, adjusting parameter weights and hyper-parameters for best prediction results. The final model was then evaluated on the test set to provide the final performance metrics. Sensitivity ("recall"), specificity, positive predictive value (PPV or "precision"), negative predictive value (NPV), and accuracy of the final model were calculated. A more detailed discussion of the three stages, including the importance of such a setup, can be found in Appendix A.6.

3. Results

The study population contained 514 patients, with 227 positive findings on CT, and 287 negative findings. Table 1 gives a breakdown of the proportions of gender, dementia presence, aspirin usage, injury above the clavicle, and fall mechanisms, broken down by positive and negative CT findings. Table 2 shows the final performance on the cross validation dataset, while Table 3 shows the final performance on the test dataset. Additionally, Tables 4 & 5 show the 2x2 contingency tables for the cross validation and test datasets, respectively.

Feature	CT+	CT−
Gender	75% Female	66% Female
Dementia	33.8%	44.2%
Aspirin Use	54.5%	46.9%
Injury above clavicle	75.0%	64.9%
Fall from bed	15.9%	9.9%
Fall from sitting	9.0%	17.1%
Fall from standing	63.6%	66.8%
Fall from Height	9.0%	5.2%

Table 1: Study Population Aggregated Statistics

Metric	Value (95% CI)
Sensitivity (Recall)	100.00% (92.13%-100.00%)
Specificity	78.95% (66.11%-88.62%)
PPV (Precision)	78.95% (66.11%-88.62%)
NPV	100.00% (92.13%-100.00%)
Accuracy	88.24% (80.35%-93.77%)

Table 2: Final Model Performance on Cross Validation Data

Metric	Value (95% CI)
Sensitivity (Recall)	97.78% (88.23%-99.94%)
Specificity	89.47% (78.48%-96.04%)
PPV (Precision)	88.00% (75.69%-95.47%)
NPV	98.08% (89.74%-99.95%)
Accuracy	93.14% (86.37%-97.20%)

Table 3: Final Model Performance on Test Data

	Predicted Positive	Predicted Negative
Actual Positive	45	0
Actual Negative	12	45

Table 4: 2x2 Contingency Table on Cross Validation Data

	Predicted Positive	Predicted Negative
Actual Positive	44	1
Actual Negative	6	51

Table 5: 2x2 Contingency Table on Test Data

4. Discussion

Existing clinical decision rules for minor head injury use age over 65 years as a predictor of high risk for intracranial injury but do not differentiate level of risk among those patients over age 65, despite only ∼2% of these patients having any acute findings. Our results using a trained ANN model compare favorably or exceed those of the studies used to establish the decision strategies that are currently being used to evaluate head trauma patients. Granted, the reader must keep in mind that all previous studies involve a general adult study population not limited by age $\geq$ 65 years or by mechanism of injury (falls in our case). Additionally, this study represents a preliminary foray into the application of ANNs to this problem, and larger cohorts with prospective, possibly multi-institutional, data collection will be needed to validate and refine the approach.

4.1. Significance

ANNs are powerful models in that they aim to automatically learn complex, nonlinear relationships between variables, such as those found in complex biological systems. In our situation, we have effectively taken a clinical problem for which an effective solution has not been found, and developed an ANN model with relatively good results as a good first step study.

4.2. Regarding Sensitivity, PPV, Specificity, and NPV

As a quick recall, sensitivity refers to the percentage of actual positives that are predicted correctly (the percentage of patients with positive CT findings that the model "catches"), specificity refers to the percentage of actual negatives that are predicted correctly, PPV refers to the percentage of predicted positives that are actually positive (the "precision" of the model), and NPV refers to the percentage of predicted negatives that are actually negative.

Currently in practice, *all* patients $\geq$ 65 years old presenting with minor head injury after a fall receive a head CT, which can be interpreted as "predicting" that all of these patients are positive for acute findings. Therefore, current practice is 100% sensitive for detecting intracranial injuries in this patient population, yet is 0% specific. Furthermore, given that only ∼2% of these patients will have acute findings, as noted in section 1, and all of these patients are currently predicted to be positive, current practice has a PPV of only ∼2%. In other words, current practice is only correct (precise) ∼2% of the time. Since no patients are currently "predicted" to be negative, NPV is undefined.

The goal in building a prediction model is to be both sensitive and precise. Maintaining high sensitivity is important in order to not miss positive patients, and the addition of high PPV (precision), and in turn high specificity, is important for the model to become a powerful part of clinical care. Reynolds [27] eloquently stated in 2013 that "what we need is an analytic framework for integrating regional culture into decisions about where to draw the 'relevance border' around a given study's conclusions", in commenting upon the difference between the multiple studies investigating the performance of the various decision rules in North America, Australia, Netherlands, and Spain. Reynolds further noted that the specificity of the New Orleans Criteria at 3-5% eliminates its utility as a mechanism to reduce CT usage in the mild traumatic brain injury population. Constructing a model with ∼100% sensitivity, but only ∼0% PPV and ∼0% specificity does not provide benefit over simply "predicting" that *all* patients are positive, such as is currently done in our clinical scenario for the given subset of patients. Therefore, our goal was to build a model with both high sensitivity *and* high PPV (precision); cases in which these are both high will in turn *also* have high specificity, a metric that in our experience is reported more often than PPV.

4.3. Comparison to Previous Studies

Our artificial neural network compared favorably to previous studies. Of note, we were unable to reconstruct the 2 X 2 table necessary to calculate sensitivity, specificity, PPV, NPV, and accuracy in all cases *except* the work of Bouida W, et al. published in 2013 [28], and Sun Ro and colleagues published in 2011 [29]. All other studies [6, 7, 8, 9] require the reader to accept the sensitivity and specificity as reported by the authors. Please see Table 6 for a listing of the numerous studies using risk stratification to identify those patients needing head CT in a general population with minor head injury mechanism.

Morton and Korley [30] discussed aspects of the CCHR and NOC, as they have been shown to be the most sensitive and specific at identifying patients with clinically important intracranial lesions in those patients with mild traumatic brain injury. In an attempt to clarify for residents making decisions, they noted that the NOC uses seven criteria, and if all seven are absent, then a CT scan is not warranted for the patient; conversely, the presence of one or more factors triggers the need for a CT scan. Using this rule, the NOC had a sensitivity of 100% for the detection of an intracranial injury, with a specificity of 24%. They noted the CCHR to have 5 high risk factors

and 2 medium risk factors. The presence of one or more of the 5 high risk factors was 100% sensitive and 68.7% specific for predicting the need for neurosurgical intervention, and the presence of one or more of the seven criteria was 98.4% sensitive and 49.6% specific for predicting clinically important brain injury. The artificial neural network performed similarly to all listed in Table 6 in terms of sensitivity (97.8%), and greatly exceeded all in terms of specificity (89.5%).

Reference	Patients	Sensitivity	Specificity
NEXUS II [7]	13,728		
Clinically Important	917	98.30%	13.70%
Minor	330	95.20%	17.30%
Bouida, et al [28]	1,582		
Clinically Important 13.8%			
CCHR		95%	65%
NOC		86%	28%
Neurosurgical Intervention 2.1%			
CCHR		100%	100%
NOC		82%	99%
Sun Ro, et al [29]	7,131		
Clinically Important 9.7%			
CCHR		79.20%	41.30%
NEXUS II		88.70%	46.50%
NOC		91.90%	22.40%
Neurosurgical Intervention 2%			
CCHR		100%	38.20%
NEXUS II		95.10%	41.40%
NOC		100%	20.40%
any traumatic finding			
CCHR		77.80%	41.90%
NEXUS II		84.90%	46.20%
NOC		91.10%	22.90%
Stiell, et al [13]	2,707		
Clinically Important 8.5%			
CCHR		100%	76.30%
NOC		100%	12.10%
Neurosurgical Intervention 1.5%			
CCHR		100%	76.30%
NOC		100%	12.10%
Smits, et al [9]	3,181		
Clinically Important 9.8%			
CCHR		83.40%	39.70%
NOC		97.70%	5.60%
Neurosurgical Intervention 0.5%			
CCHR		100%	37.20%
NOC		100%	3.00%
Wolf, et al [31]	12,786		
CT Received	1,307		
Intracranial Injury	489		
Novel Criteria		90%	67%
CCHR		80%	72%
Stiell, et al [6]	3,121		
Clinically Important 8%			
CCHR medium risk		98.40%	49.60%
Neurosurgical Intervention 1%			
CCHR high risk		100%	68.70%
Haydel, et al [8]			
NOC phase 1	520	94%	
NOC phase 2	909	100%	25%
Stein, et al [14]	7,955		
CCHR High Risk		97%	51%
CCHR Medium Risk		99%	47%

(Continued on next page)

National Institute of Clinical Excellence		99%	31%
Neurotraumatology Committee		96%	47%
NEXUS II		97%	47%
NOC		99%	33%
Scandinavian		96%	53%
Korley and Morton [32]	169		
CT+	5		
2008 ACEP Guidelines		80%	10.40%
CCHR		100%	36.80%
NOC		100%	3.20%

Canadian CT Head Rule (CCHR)
National Emergency X Ray Utilization Study II (NEXUS II)
New Orleans Criteria (NOC)

Table 6: Summary of Previous Studies

4.4. Clinical Usage

Regression models are quite useful for modeling linear relationships, and have the ability to identify important features that correlate to the final diagnosis. These important features can then be used directly to form decision rules for clinical practice, such as the case with the CCHR. In contrast, ANNs lack the ability to provide explicit correlations [33], but can be more effective for complex scenarios with nonlinear relationships. Therefore, clinical workflows with ANNs would involve direct use of the trained ANN models, in which clinical features for a given patient would be fed into the model, and a prediction would be produced. The prediction could then be used by the provider as an aid in making clinical decisions during the care of the patient, just as decision rules are used. We believe that this is a worthwhile tradeoff for increased objective data, as the ultimate goal is to provide the best possible patient care.

Actual use of the models in clinical workflows does not have to be difficult though. The models could be integrated into the electronic health record (EHR) so that as the provider completed the chart during a patient encounter, the relevant clinical features would feed into a built-in ANN model that has already been trained. Upon logging all of the needed clinical features, the built-in ANN would output the probability/prediction of acute CT findings being present (for this clinical scenario), allowing the provider to make a better-informed clinical decision about ordering a CT.

5. Limitations

We note again that the population used in this study is different from, and a subset of, the general populations studied in the papers discussed in section 4.3. We specifically studied the population of patients ≥ 65 years of age presenting with minor head injury after a fall, since no rules have been formulated specifically for stratifying this class of patient. Regardless, the discussions in section 4.3 are important as they allow the reader to compare the results of our study with the results and currently accepted practices in other age groups for the similar scenario.

Additionally, the data used to "train", "cross validate", and "test" the artificial neural network is retrospective and has the limitations commonly associated with such data. Our cohort was defined by those patients in the studied population who received a head CT, and thus does not include those who did not undergo a CT scan and may have potentially had missed injuries. However, the current state of integration of electronic medical records and artificial neural networks is essentially nil and precludes the collection of prospective data in this regard.

Finally, while this study represents a good first step, larger retrospective and prospective cohorts, possibly with multi-institutional data collection, will be needed to narrow the confidence intervals, refine the performance point estimates, and validate the overall approach.

6. Conclusion

Artificial neural networks show great potential for predicting, with high sensitivity and high specificity, the need for head CTs in patients age ≥ 65 presenting with minor head injury after a fall. As a preliminary foray into this application, our ANN showed comparable sensitivity, predictive values, and accuracy, with a much higher specificity than the existing decision rules in clinical usage for predicting head CTs with acute intracranial findings. We believe that with growing amounts of medical data available from patients, and a need to predict results for increasingly complex cases, the use of ANNs may become increasingly effective. In addition to validating our approach with larger studies, future goals could include effectively integrating ANNs into clinical workflows, and exploring the application of ANNs to other complex clinical cases.

Appendix A. Artificial Neural Networks

The following discusses ANNs in more depth, with specific regard to the type and formulation used in this study, and uses information gathered from various sources [18, 21, 22, 23]. The interested reader is encouraged to consult this appendix and the related sources for further information.

Appendix A.1. Biological Neurons

The concept and motivation for ANNs has connections to neuroscience. As a high-level overview, neurons in the human brain receive inputs via their dendrites, and emit an output through a single axon. Connections between dendrites and axons can be thought of as having varying "strengths", which provide for a "weighted" set of inputs. The input pulses converge on the body of the neuron, and if the "weighted summation" of energy in the body exceeds a certain threshold, the neuron will fire. The combination of many such neurons forms a biological neural network.

Appendix A.2. The Artificial Neuron

An artificial neural network is an algorithmic abstraction of these concepts largely developed in the fields of computer science and mathematics. In the algorithmic version, biological neurons are simulated at a basic, highly abstracted level: each node accepts numeric inputs; assigns a numeric weight to each input feature; calculates a weighted summation (linear combination) of the inputs; adds a "bias" term (which can intuitively be thought of as the threshold value); applies a nonlinear function to this value; and then outputs the result. The output a, or "activation", of a neuron for a single example can be calculated as

$$a = g\left(\left(\sum_{i=1}^{n} x_i w_i\right) + b\right), \tag{A.1}$$

or more concisely, using linear algebra notation, as

$$a = g(x^T w + b), \tag{A.2}$$

where x is a vector of n input features for a single example, w is a vector of n weights, b is a scalar threshold ("bias"), and g is a nonlinear function. For hidden layer nodes, we use the rectified linear unit (ReLU) nonlinear function

$$g(z) = \max(0, z), \tag{A.3}$$

which thresholds an input z at 0. For the output node, we use the logistic, or "sigmoid", nonlinear function

$$g(z) = \frac{1}{1 + e^{-z}}, \tag{A.4}$$

which maps an input z to a value between 0 and 1, allowing for a probabilistic interpretation.

Overall, the power of an ANN arises from the nested combination of many neurons into one overall model.

Appendix A.3. Layers of Artificial Neurons

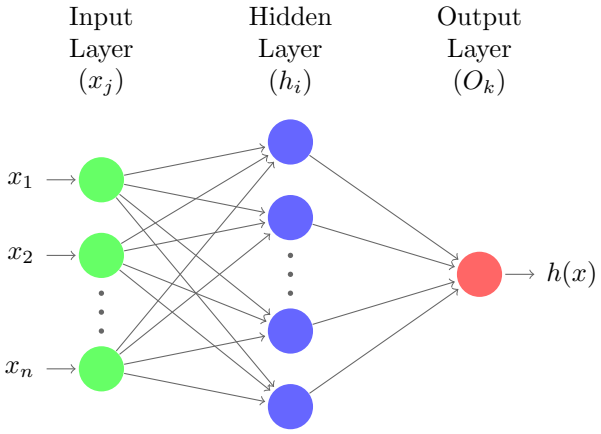

Figure A.1: Artificial Neural Network Diagram

Individual neurons are grouped into successive "layers" to form a network, with a traditional layout beginning with an "input" layer, followed by one or more "hidden" layers, and ending with a final "output" layer. Figure A.1 shows the relationship of the neurons in a typical example network. The input layer consists of a set of pseudo-neurons, with one for each input feature in the dataset; these simply output the value of that input, unaltered, and therefore can simply be thought of as supplying the input data to the rest of the network. The hidden layer is a layer of a varying number of neurons (the exact number is decided upon while training), where each neuron receives and processes all of the outputs of the input layer (which are just the unaltered inputs) using equation A.2 & A.3. The output layer consists of k neurons equal to the number of target outputs given in a vector y for each example in the dataset, with each neuron receiving and processing all of the outputs of the hidden layer using equations A.2 & A.4. For a dataset with a single "true"/"false" answer for each example (such as the presence of an acute CT finding represented in the dataset by a single y value equal to 1 or 0), there would be one neuron in the output layer, with the output a equal to the probability that the answer is "true". As an alternative example, for a dataset where the answer for each example could be one of k different possibilities, or "classes" (such as k different diseases), there would be k neurons in the output layer, each outputting the probability of its respective class (respective disease) being "true".

The hypothesis $h(x)$ of the overall network,

$$h(x) = a_{\text{outputs}}, \tag{A.5}$$

is equal to a vector a_{outputs} consisting of the output values of the output neurons, with each value interpreted as the probability of the respective answer being "true" (a value of 0 would suggest that the answer is "false").

Appendix A.4. Flow of Information

In an ANN, information flows one layer at a time towards the output neurons, performing all of the processing for a given layer before moving on to the next layer. For a single example (single patient in our case), the input neurons are provided a vector x of n features for the example (single patient), and an output hypothesis vector is obtained once the processing is complete. For an entire dataset, this process is effectively repeated for each example, although in practice it will be done in parallel by making use of matrices.

Appendix A.5. Learning

In order for the ANNs to learn how to correctly predict results, the "θ parameters" (weights and bias collectively) for each neuron must be adjusted in a "training" step.

To begin, the network develops a hypothesis vector for each example (each patient) in the "training set" based on the associated input features. The hypothesis for the i^{th} example (i^{th} patient) is then compared to the example's target outcome y to determine the measure of fit ("loss" or "cost") for that example.

Given our ANN architecture, the "loss" or "cost" for an example i can be computed as

$$L^{(i)}_{\text{data}} = -y^{(i)^T} \log(h(x^{(i)})) - (1 - y^{(i)})^T \log(1 - h(x^{(i)})), \tag{A.6}$$

where $y^{(i)}$ is the target outcome vector for the ith example (patient), and $L^{(i)}_{\text{data}}$ is a scalar value that is the negative log likelihood assuming a Bernoulli distribution. Essentially, this loss is a measure of how well the hypothesis $h(x^{(i)})$ matches the target $y^{(i)}$ for the i^{th} example.

In addition to the loss for the data, we also introduce an "L2 regularization" term,

$$L_{\text{reg}} = \frac{\lambda}{2} \sum [(w^{(l)}_{ij})^2], \tag{A.7}$$

that sums all of the squared weights in the network, where l equals the layer number, ij refers to the connection between the j^{th} neuron in layer l and the i^{th} neuron in layer

$l+1$, λ ("*lambda*") equals a hyper-parameter that adjusts the level of regularization, and L_{reg} is a scalar value. This regularization term places a zero-mean Gaussian prior on the weights, and is used to promote generalizability of the network and prevent "overfitting" by incurring a high loss for large weight values.

The overall loss of the ANN for the entire dataset is then computed as

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} L_{\text{data}}^{(i)} + L_{\text{reg}}, \tag{A.8}$$

where the first term is the average data loss over m examples, and the latter is the regularization term. A high loss L indicates that the model is performing poorly (due to large prediction errors for the examples, and/or large weight values reducing generalizability).

The final step is to minimize the loss L for the training set of patients by adjusting the network θ parameters (weights & biases). For this step, we use a method known as "minibatch gradient descent". This step is performed by mathematically computing the partial derivatives of the loss L with respect to the individual θ parameters, and then adjusting each of the parameters by a small fraction of their respective partial derivative in the direction that minimizes L. The simplistic version of this update can be stated as

$$\theta_{ij}^{(l)} := \theta_{ij}^{(l)} - \alpha\frac{\partial L(\theta)}{\partial \theta_{ij}^{(l)}}, \tag{A.9}$$

where l equals the layer number, ij refers to the connection between the j^{th} neuron in layer l and the i^{th} neuron in layer $l+1$, and α equals a "learning rate" hyper-parameter that controls the magnitude of parameter updates. The partial derivatives can intuitively be thought of as the contribution (magnitude & direction) of each of the θ parameters to the loss L, with the direction pointing towards increasing values of L. Thus, by adjusting the θ parameters in the opposite direction, the loss L can be reduced, increasing the performance of the model. Note that in practice, we use a slightly different formulation known as *Adam* [34], which has better convergence properties.

These training steps are repeated for the training dataset until the loss L is sufficiently minimized, which we discuss further in section Appendix A.6.

Appendix A.6. Regarding Training, Cross Validation, and Testing

The following discusses each stage of the procedure in more detail, including the importance of such a setup.

Appendix A.6.1. Training

In this step, the neural network is trained to learn to correctly predict the results of the "training" set of patients by adjusting weights and bias values of the model per Appendix A.5. We cycle through the list of training patients repeatedly, performing learning and evaluation of error each cycle, until the average loss L across all of the training patients begins to decrease only minimally. Intuitively, this step is analogous to a student studying notes and self-quizzing. In this stage we also adjust various hyper-parameters (number of hidden layers & neurons, cycles of training, regularization, data inputs, etc.) to create different possible models. The models with sufficiently low training error can be seen as the different possible "hypotheses" for the given problem. Since the models are attempting to learn from the data in this step, a sufficiently large portion of the overall data is needed; heuristically, 60% is generally considered an effective amount.

Appendix A.6.2. Cross validation

In this step, the various training models ("hypotheses") are "validated" on a separate, smaller "cross validation" (CV) set of data to select the best model, which will be the one with the best performance on this CV set. No learning occurs during the CV stage, and the models are simply evaluated for performance; this is analogous to a student repeatedly taking a practice test in which only the grade is given, adjusting study techniques in between. This step ensures that the neural network is learning the (complex) relationships between input features during training, without simply memorizing the training examples (training patients), and is accomplished by selecting the best hyper-parameters to allow for such learning. Learning feature relationships is key to the goal of the model being able to predict results for new patients in clinical scenarios; just being able to determine the correct answer for past patients is not useful clinically. As there may be multiple hypothesis models that have low training error, this step is necessary to select the best model, which is the one that generalizes to this new set the best. Furthermore, we often will repeat the training and CV steps multiple times in order to ultimately select the best combination of hyper-parameters. Since this step is used to validate models, rather than teach them, a smaller portion of data is needed than for training; 20% is generally an effective amount.

Appendix A.6.3. Testing

Finally, once we feel that our model has actually "learned" and has been selected for maximum performance on the CV data, we evaluate it on another separate, small set of "test" patients and report the results on this test set as the final performance of our model, as seen in section 3. This step effectively allows the study to be replicated within the study itself, as this test set of patients is not used by the model during the rest of the training process, and therefore acts as a completely new set of patients. Since the algorithm selects for the model with the best performance on the CV set of data, the model is implicitly tied to the CV set. In contrast, the test set is only used to evaluate the final model, and is never used as a means of selecting models. Therefore, the test set of data is not tied to the model, and thus can unbiasedly determine the model's future effectiveness in clinical scenarios; this is analogous to

a student taking a final exam. As with the CV set, using 20% of the original data is a generally effective heuristic.

Appendix A.7. Regarding Hyper-Parameters

Using cross validation (sections 2.4 & Appendix A.6) with randomized hyper-parameter search [35], a hypothesis model with the settings & hyper-parameters listed in Table A.7 was selected as the final model. Hyper-parameters included: number of hidden layers; number of hidden neurons per hidden layer; λ value used in regularization; and a "probability threshold" with which to interpret the hypothesis of the network. This probability threshold is not to be confused with the intuitive idea of a neuron threshold, and is used in this context to interpret the hypothesis values (which are probabilities between 0 and 1) as equating to either "true" or "false". Values above the threshold value are considered "true". The randomized hyper-parameter search allowed us to test large ranges of possible values while quickly narrowing down on ideal ranges, leading to the final combination.

Setting/Hyper-parameter	Value
Num Input Neurons	8
Num Hidden Layers	1
Num Hidden Neurons/Layer	25
Num Output Neurons	1
λ (Regularization)	0.001
Probability Threshold	0.18

Table A.7: Final Model Settings & Hyper-parameters

Appendix A.8. Regarding Skewed Classes

Rare conditions are common in the medical world, however, attempting to model these situations in which there will be a large class imbalance if data is gathered completely randomly can lead to difficulties in learning, as noted by Mazurowski, et al. [25]. If a rare condition (such as that presented in this paper) is only present in a limited number of examples in the dataset (corresponding to the natural prevalence of that condition), then a model that tends to predict the "negative condition" for all examples may seem to have high performance as measured by low test error, 100% specificity, and high accuracy. However, this model would have a sensitivity of 0%, and thus would not be effective for clinical (or any other kind of) usage.

Models such as ANNs or logistic regression models need to learn the characteristics and feature relationships of both positive and negative cases in order to create a "decision boundary" that can effectively separate the two possible conditions [18]. The prevalence of the condition is not used with these types of models. If there is only a limited number of examples of positive cases as compared to negative cases (and vice versa), it becomes much more difficult for the models to learn relationships, and thus performance will be impacted. Therefore, an effective strategy is choice-based sampling [24] in which roughly equal numbers of both positive and negative cases are gathered. Learning an effective decision boundary allows the model to properly distinguish between positive and negative cases, thus allowing it to achieve both high sensitivity and high specificity.

Appendix A.9. Regarding Feature Correlation

Logistic regression models determine coefficients for each input feature, calculate a weighted sum from these coefficients and an example set of inputs, and then directly transform this sum to a value between 0 and 1 to output a probability of the answer being true. In these logistic regression models, since the coefficients are linearly related to the output, the features corresponding to the coefficients with the greatest magnitudes can be interpreted as being correlated to the final diagnosis. Thus, if we are able to model a problem with linear methods, then we can use the important features directly as clinical markers to form decision rules.

ANNs compute several nonlinear transformations of the original inputs before outputting a result. Thus, the coefficients for the input features are not linearly related to the outcome, and cannot be interpreted as correlations to important features. Therefore, effective clinical usage of ANNs would involve the use of the model itself to form predictions, as discussed in section 4.4, rather than an interpretation of correlated clinical markers forming decision rules.

Acknowledgements & Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] G. Bergen, L. Chen, M. Warner, L. Fingerhut, Injury in the United States: 2007 Chartbook., National Center for Health Statistics, Hyattsville, MD, 2008.

[2] Centers for Disease Control, WISQARS (Web-based Injury Statistics Query and Reporting System).
URL http://cdc.gov/injury/wisqars/

[3] A. S. Gangavati, D. K. Kiely, L. K. Kulchycki, R. E. Wolfe, J. L. Mottley, S. P. Kelly, L. A. Nathanson, A. P. Abrams, L. A. Lipsitz, Prevalence and characteristics of traumatic intracranial hemorrhage in elderly fallers presenting to the emergency department without focal findings, Journal of the American Geriatrics Society 57 (8) (2009) 1470–1474. doi:10.1111/j.1532-5415.2009.02344.x.

[4] D. R. Martin, R. C. Semelka, Health effects of ionising radiation from diagnostic CT, The Lancet 367 (9524) (2006) 1712–1714. doi:10.1016/S0140-6736(06)68748-5.

[5] D. J. Brenner, E. J. Hall, Computed tomography—an increasing source of radiation exposure, New England Journal of Medicine 357 (22) (2007) 2277–2284. doi:10.1056/NEJMra072149.

[6] I. G. Stiell, G. A. Wells, K. Vandemheen, C. Clement, H. Lesiuk, A. Laupacis, R. D. McKnight, R. Verbeek, R. Brison, D. Cass, M. E. Eisenhauer, G. Greenberg, J. Worthington, The Canadian CT Head Rule for patients with minor head injury., The Lancet 357 (9266) (2001) 1391–1396. doi:10.1016/S0140-6736(00)04561-X.

[7] W. R. Mower, J. R. Hoffman, M. Herbert, A. B. Wolfson, C. V. Pollack Jr, M. I. Zucker, N. I. Investigators, et al., Developing a decision instrument to guide computed tomographic imaging of blunt head injury patients, Journal of Trauma and Acute Care Surgery 59 (4) (2005) 954–959. doi:10.1097/01.ta.0000187813.79047.42.

[8] M. J. Haydel, C. A. Preston, T. J. Mills, S. Luber, E. Blaudeau, P. M. DeBlieux, Indications for computed tomography in patients with minor head injury, New England Journal of Medicine 343 (2) (2000) 100–105. doi:10.1056/NEJM200007133430204.

[9] M. Smits, D. W. Dippel, G. G. de Haan, H. M. Dekker, P. E. Vos, D. R. Kool, P. J. Nederkoorn, P. A. Hofman, A. Twijnstra, H. L. Tanghe, et al., External validation of the Canadian CT Head Rule and the New Orleans Criteria for CT scanning in patients with minor head injury, JAMA 294 (12) (2005) 1519–1525. doi:10.1001/jama.294.12.1519.

[10] F. Servadei, G. Teasdale, G. Merry, Defining acute mild head injury in adults: a proposal based on prognostic factors, diagnosis, and management, Journal of Neurotrauma 18 (7) (2001) 657–664. doi:10.1089/089771501750357609.

[11] National Collaborating Centre for Acute Care (UK and others), Head injury: triage, assessment, investigation and early management of head injury in infants, children and adults, National Collaborating Centre for Acute Care (UK), 2007.
URL http://www.ncbi.nlm.nih.gov/pubmed/21563330

[12] T. Ingebrigtsen, B. Romner, C. Kock-Jensen, Scandinavian guidelines for initial management of minimal, mild, and moderate head injuries, Journal of Trauma-Injury, Infection, and Critical Care 48 (4) (2000) 760–766.
URL http://www.ncbi.nlm.nih.gov/pubmed/10780615

[13] I. G. Stiell, C. M. Clement, B. H. Rowe, M. J. Schull, R. Brison, D. Cass, M. A. Eisenhauer, R. D. McKnight, G. Bandiera, B. Holroyd, et al., Comparison of the Canadian CT Head Rule and the New Orleans Criteria in patients with minor head injury, JAMA 294 (12) (2005) 1511–1518. doi:10.1001/jama.294.12.1511.

[14] S. C. Stein, A. Fabbri, F. Servadei, H. A. Glick, A critical comparison of clinical decision instruments for computed tomographic scanning in mild closed traumatic brain injury in adolescents and adults, Annals of Emergency Medicine 53 (2) (2009) 180–188. doi:10.1016/j.annemergmed.2008.01.002.

[15] K. Ono, K. Wada, T. Takahara, T. Shirotani, Indications for computed tomography in patients with mild head injury, Neurologia medico-chirurgica 47 (7) (2007) 291–298. doi:10.2176/nmc.47.291.

[16] J. M. Bennett, N. R. Nehus, M. R. Astin, C. K. Brown, R. Johnson, K. L. Brewer, Use of cranial computed tomography (CT) in elderly patients presenting after a fall: Can we predict those having abnormal head CT scans, British Journal of Medicine & Medical Research 6 (3) (2014) 342–350. doi:10.9734/BJMMR/2015/10435.

[17] A. Riccardi, F. Frumento, G. Guiddo, M. B. Spinola, L. Corti, P. Minuto, R. Lerza, Minor head injury in the elderly at very low risk: a retrospective study of 6 years in an emergency department (ED), The American Journal of Emergency Medicine 31 (1) (2014) 37–41. doi:10.1016/j.ajem.2012.05.023.

[18] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, Vol. 2, Springer, 2009.

[19] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks 4 (2) (1991) 251–257. doi:10.1016/0893-6080(91)90009-T.

[20] M. Sinha, C. S. Kennedy, M. L. Ramundo, Artificial neural network predicts CT scan abnormalities in pediatric patients with closed head injury, Journal of Trauma and Acute Care Surgery 50 (2) (2001) 308–312. doi:10.1097/00005373-200102000-00018.

[21] D. J. MacKay, Information theory, inference, and learning algorithms, Vol. 7, Citeseer, 2003.

[22] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[23] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, book in preparation for MIT Press (2016).
URL http://www.deeplearningbook.org

[24] J. Hosek, An introduction to estimation with choice-based sample data.
URL http://www.rand.org/pubs/papers/P6361.html

[25] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, G. D. Tourassi, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural Networks 21 (2) (2008) 427–436. doi:10.1016/j.neunet.2007.12.031.

[26] F. Chollet, Keras, https://github.com/fchollet/keras (2015).

[27] T. A. Reynolds, A Tunisian, a Canadian, and an American walk into a bar (sustaining mild head injury), Annals of Emergency Medicine 61 (5) (2013) 528–529. doi:10.1016/j.annemergmed.2012.12.006.

[28] W. Bouida, S. Marghli, S. Souissi, H. Ksibi, M. Methammem, H. Haguiga, S. Khedher, H. Boubaker, K. Beltaief, M. H. Grissa, et al., Prediction value of the Canadian CT Head Rule and the New Orleans Criteria for positive head CT scan and acute neurosurgical procedures in minor head trauma: a multicenter external validation study, Annals of Emergency Medicine 61 (5) (2013) 521–527. doi:10.1016/j.annemergmed.2012.07.016.

[29] Y. S. Ro, S. D. Shin, J. F. Holmes, K. J. Song, J. O. Park, J. S. Cho, S. C. Lee, S. C. Kim, K. J. Hong, C. B. Park, et al., Comparison of clinical performance of cranial computed tomography rules in patients with minor head injury: a multicenter prospective study, Academic Emergency Medicine 18 (6) (2011) 597–604. doi:10.1111/j.1553-2712.2011.01094.x.

[30] M. J. Morton, F. K. Korley, Head computed tomography use in the emergency department for mild traumatic brain injury: integrating evidence into practice for the resident physician, Annals of Emergency Medicine 60 (3) (2012) 361–367. doi:10.1016/j.annemergmed.2011.12.026.

[31] H. Wolf, W. Machold, S. Frantal, M. Kecht, G. Pajenda, J. Leitgeb, H. Widhalm, S. Hajdu, K. Sarahrudi, Risk factors indicating the need for cranial CT scans in elderly patients with head trauma: an Austrian trial and comparison with the Canadian CT Head Rule: clinical article, Journal of Neurosurgery 120 (2) (2014) 447–452.

[32] F. K. Korley, M. J. Morton, P. M. Hill, T. Mundangepfupfu, T. Zhou, A. M. Mohareb, R. E. Rothman, Agreement between routine emergency department care and clinical decision support recommended care in patients evaluated for mild traumatic brain injury, Academic Emergency Medicine 20 (5) (2013) 463–469. doi:10.1111/acem.12136.

[33] J. V. Tu, Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, Journal of clinical epidemiology 49 (11) (1996) 1225–1231. doi:10.1016/S0895-4356(96)00002-9.

[34] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.
URL http://arxiv.org/abs/1412.6980

[35] Y. Bengio, Practical recommendations for gradient-based training of deep architectures, CoRR abs/1206.5533.
URL http://arxiv.org/abs/1206.5533