Large Synoptic Survey Telescope

www.lsst.org

Petascale Object Classification of the LSST Event Stream

K. D. Borne (GMU), R. Laher (Caltech), Z. Ivezic (UW), N. Hamam (Caltech), and the LSST Collaboration

We describe a specific example of object classification using ANN (Artificial Neural Networks) applied to data from the SDSS (Sloan Digital Sky Survey). We test several ANN models for classification *Reliability (R)* and *Completeness (C)*. The ANN performs better than a deterministic classifier by a factor of 3 in the *Unreliability (1-R)* metric. These experiments are a precursor to the large-scale object classification that will be required for the LSST. The LSST object database will contain detailed information for 20 billion sources, including approximately 10 billion galaxies, a similar number of stars, and over 1 billion variable sources (optical variables, transients, or moving objects). After 10 years of survey operations, the LSST object database will comprise 10-20 Petabytes of science catalog attributes: over 200 science attributes per object will be available for classification, characterization, and mining. Deep multi-color photometry and long-term time series (photometric and astrometric, at various cadences) will yield an enormously rich potential for new scientific discoveries. LSST's impressive panoply of precision parameter data will enable the characterization and classification of astronomical objects on a grand scale. Event characterization and classification within seconds of each exposure will permit timely knowledge-based follow-up on the most significant and exciting astronomical discoveries of the coming decade.

Confronting the Enormous LSST Data Volume:

•The LSST sky survey will yield unprecedented insight into the astrophysical properties of >20 billion sources.
•Automated classification methods on the 20-Petatbye LSST object database will be required in order to extract all of the knowledge hidden within such an enormous data collection.
•Application of sophisticated machine learning methods that have the highest performance in terms of completeness (C) and reliability (R) will be the key to maximizing the scientific return from LSST.

Object Classification through Artificial Neural Networks (ANN):

•We explore the performance of classification methods based on artificial neural networks (ANNs).
•We compare these results against a deterministic method based on known correlations in the data.
•We demonstrate that ANNs not only have superior performance, but ANNs are also capable of automatically discovering and exploiting data correlations among astronomical-object categories that are not easily found, such as making novel discoveries from conventional color-color diagrams of astronomical sources.
•Each ANN model (**FIG. 1**) is labeled a-b-c (or a-b-c-d), where "a" specifies the # of input parameters *(e.g.,* object colors), "d" specifies the # of outputs (frequently d=1, indicating that the output is the object classification), and "b" (and "c") specify the # of hidden nodes in the ANN's first (and second) hidden layer. *(Refer to a machine learning textbook for an explanation.)*

Science Use Case:

•We illustrate how ANNs can use multi-color data from a single source detection to predict whether that source is highly variable (*e.g.,* a quasar), or non-variable (*e.g.,* an H or He white dwarf).
•By training an ANN on a smaller data set where a target property is known (variability vs. non-variability in our example), ANNs can be used as precursor (initial-pass) classifiers for sorting a large number of source detections before the full dynamic behavior of the associated astronomical object can be fully characterized, over time scales of months to years.
•Initial-pass precursor classifiers may contribute significantly to the near-real-time probabilistic (*e.g.,* Bayesian) classifications that will accompany each of the 10-100,000 nightly LSST event alert notifications.

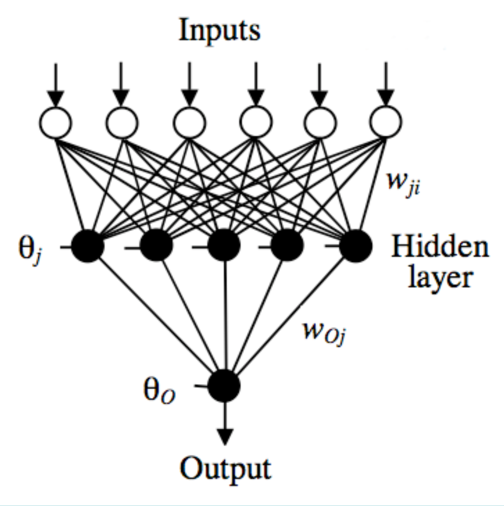


FIG. 1: Schematic of an ANN (artificial neural network) training network. The ANN effectively represents a multi-step non-linear regression. The values of the Hidden Layer nodes are a regression on the Input Layer node values (the input parameters), and the value of the Output Layer node (classification) is a regression on the Hidden nodes. The value of the Output node is usually the classification of the source data, represented by the parameters fed into the Input nodes.

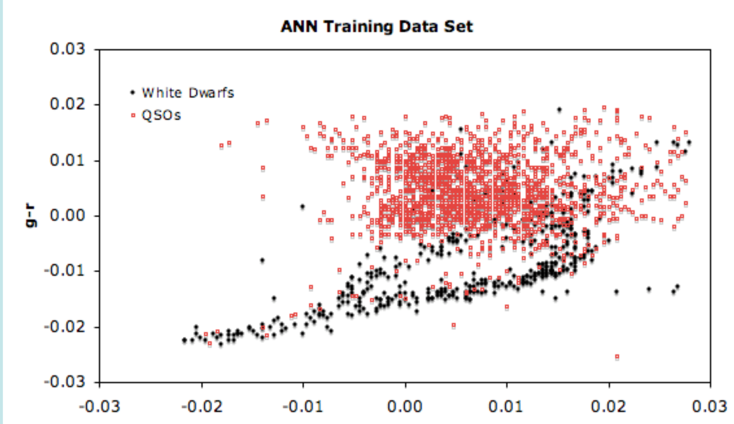


FIG. 2a: Color-color diagram of the input objects used **to train** the ANN. These are stripe82 SDSS-selected QSOs and white dwarfs, with:
320 < ra < 40
|dec| < 1.27
u < 20 (with at least 10 u obs)
-0.6 < u-g < 0.7
-0.7 < g-r < 0.5
(Ivezic *et al.* 2007, AJ, 134, 973)

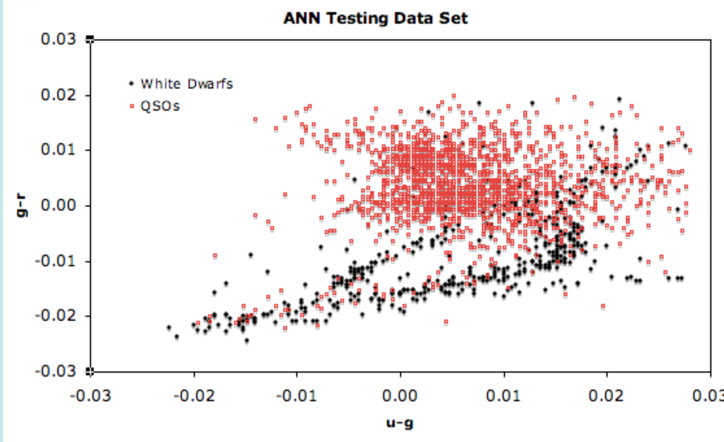


FIG. 2b: SDSS objects that were used **to test** the ANN classifier's *reliability (R)* and *completeness (C)*. These input data were selected using the same criteria as the training data, except that these data were **not** used in the ANN training.

FIG. 3a: ANN classification performance for a 6-5-5-1 network after 50.8M training iterations, with 6 input parameters: Ar, u, g, r, i, z (Ar is r-band extinction). ANN performance metrics *R=Reliability* and *C=Completeness* are plotted against ANN output threshold. The x-axis is 100x the ANN output value: QSOs correspond to x ~ 1, and White Dwarfs (WDs) have x ~ 0.

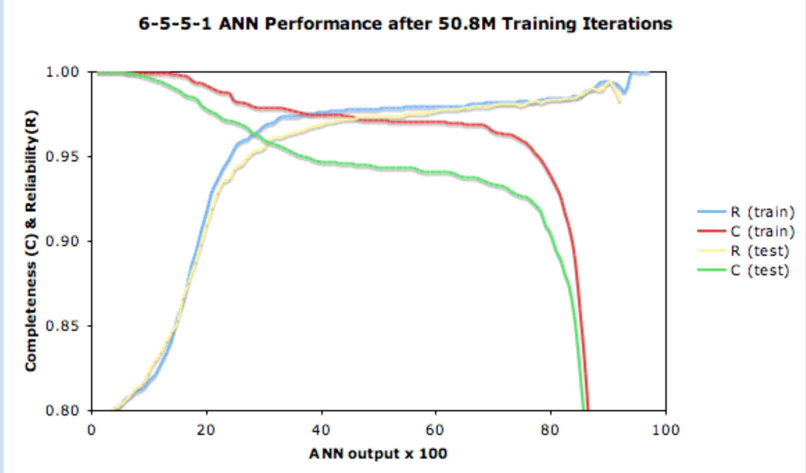
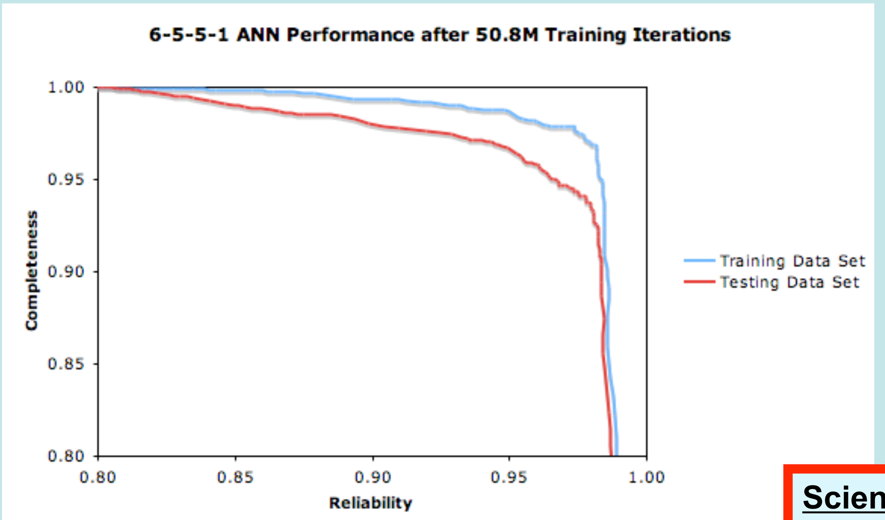
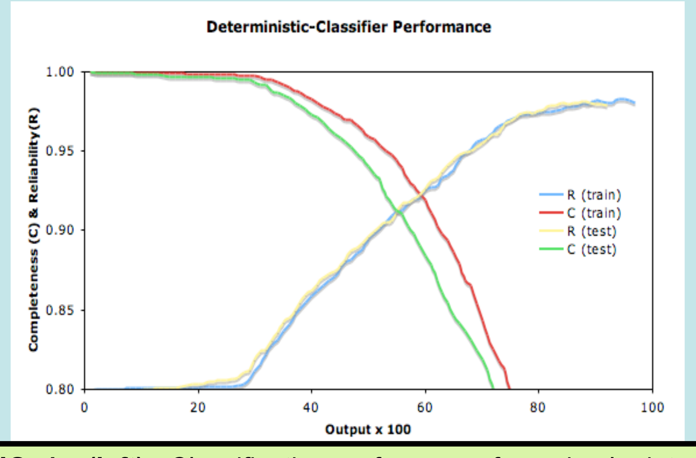


FIG. 3b: ***C* vs. *R*** discriminator curve, indicating that the optimal ANN classification performance occurs approximately where *R*=0.985 and *C*=0.95 --- **FIG.3a** indicates that this corresponds to an optimal ANN output threshold (T) approximately at x=0.80 .





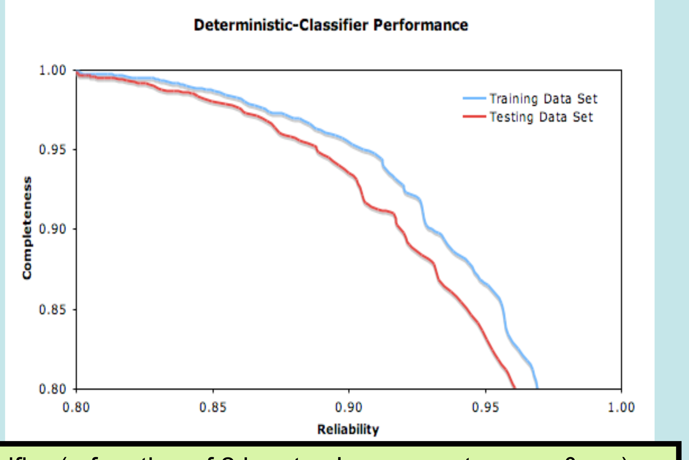
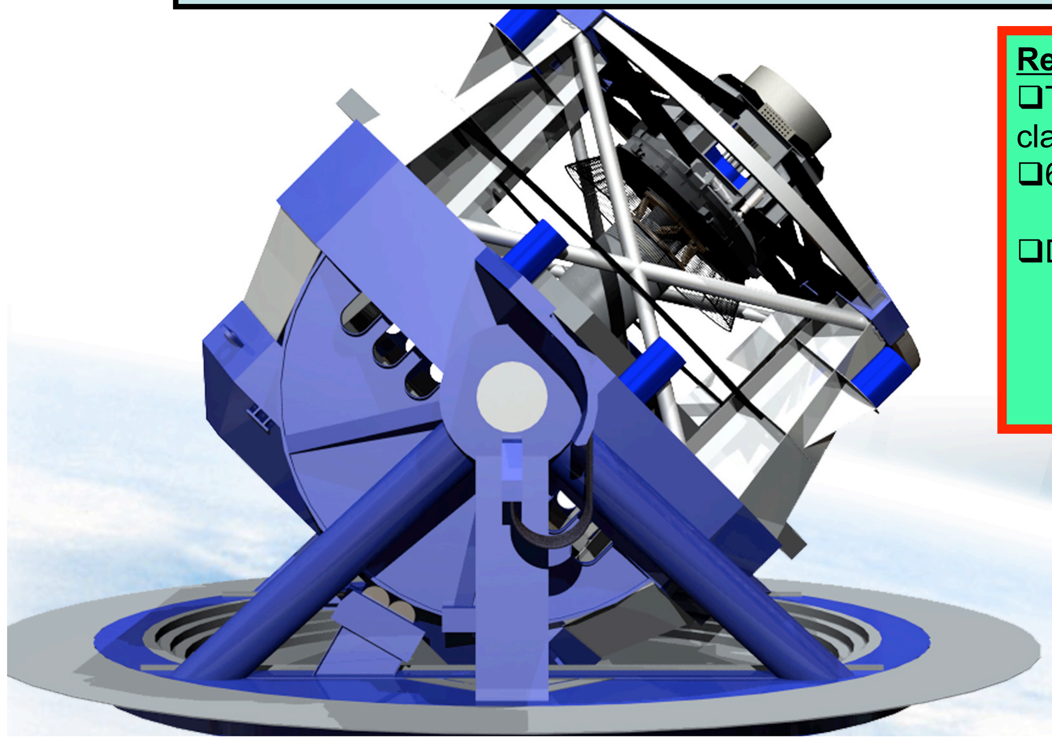


FIG. 4a (left): Classification performance for a simple deterministic classifier (a function of 2 input color parameters u-g & g-r). While the peak accuracy is competitive with the ANN, the ANN performance is superior, as indicated in **FIG.3b** vs. **FIG. 4b**.
FIG. 4b (right): ***C* vs. *R*** discriminator curve for the simple deterministic classifier of **FIG. 4a**. The accuracy of the optimal classifier (at the "knee" in the curve) is significantly lower than the accuracy of the ANN. We believe that this demonstrates that the ANN is detecting significant data correlations in the other input parameters that distinguish QSOs from White Dwarfs.

Scientific data mining use cases anticipated with the LSST database:

- Provide rapid probabilistic classifications for all 10,000 LSST events each night
- Find new "fundamental planes" of correlated astrophysical parameters
- Find new correlations, associations, relationships of all kinds from 100+ attributes in the LSST science database, integrated with distributed VO-accessible data
- Compute multi-point multi-dimensional correlation functions over the full panoply of astrophysical parameter spaces
- Discover zones of avoidance in interesting parameter spaces (*e.g.,* period gaps)
- Discover new properties of known classes
- Discover new and improved rules for classifying known classes of objects (*e.g.,* photometric-z)
- Discover new and exotic classes of astronomical objects
- Identify novel, unexpected temporal behavior in all classes of objects
- Hypothesis testing – verify existing (or generate new) astronomical hypotheses with strong statistical confidence, using millions of training samples
- Serendipity – discover rare one-in-a-billion objects through novelty detection
- Image processing – identify non-astronomical features, classify them, and separate them from the astronomical catalog inputs
- Quality assurance – identify system glitches, instrument anomalies, and pipeline errors through near-real-time deviation detection



Results of ANN experiments:

❑The ANN performs better than a deterministic classifier by a factor of 3 in the *Unreliability* metric
❑6-5-5-1 ANN classifier :
- ***C=95 R=96.5 U=3.5***

❑Deterministic classifier :
- ***C=95 R=88.9 U=11.1***
- *C = Completeness*
- *R = Reliability*
- *U = Unreliability = 1 – R*

The Utility of Artificial Neural Networks:

•ANN-discovered correlations among input parameters are hidden within the "black box" of the trained artificial neural network.
•As a consequence, ANNs thus pose an interpretation challenge: how to understand and to instantiate the discovered correlations explicitly.
•Nevertheless, ANNs can provide strong indication that correlations among parameters exist.
•By their inherent generality, ANNs are capable of simultaneously processing diverse astrometric and photometric data streams, such as spatial, extinction, and color-magnitude data.
•ANN inputs are readily configurable to handle either static or dynamic (time-series) data (or both) in the same process.

LSST is a public-private partnership. Design and development activity is supported in part by the National Science Foundation. Additional funding comes from private gifts, grants to universities, and in-kind support at Department of Energy laboratories and other LSSTC Institutional Members.