# 7. Database Preparation[5]

American Joint Committee on Cancer (AJCC) has propounded the popular Tumor-Node-Metastasis (TNM) system (Table 1) for cancer staging. A further abstraction on TNM staging is the AJCC staging which may be very helpful if used as the common class variable in a supervised machine learning method. For example, Table 2 depicts the AJCC staging derivation from the TNM staging of lung cancer. Different types of cancers have different propositions for TNM staging, but AJCC staging brings them under a common platform and if the AJCC staging is automated, the respective TNM staging may be easily determined irrespective of the cancer type.

| | | |
|---|---|---|
| Primary Tumor (T) | TX | Primary tumor cannot be assessed |
| | T0 | No evidence of primary tumor |
| | $T_{is}$ | Carcinoma in situ |
| | T1 | Tumor size <=3cm |
| | T1a | Tumor size <=2cm |
| | T1b | Tumor size >2cm and <=3cm |
| | T2 | Tumor size >3cm and <=7cm |
| | T2a | Tumor size >3cm and <=5cm |
| | T2b | Tumor size >5cm and <=7cm |
| | T3 | Tumor size >7cm |
| | T4 | Any size that invades other organs such as heart, nervous system, etc. |
| Regional Lymph Nodes (N) | NX | Regional Lymph Nodes cannot be assessed |
| | N0 | No Regional Lymph Node metastases |
| | N1 | Metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes, including involvement by direct extension |
| | N2 | Metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s) |
| | N3 | Metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s) |
| Distant Metastasis (M) | M0 | No Distant Metastasis |
| | M1 | Distant Metastasis |
| | M1a | Separate tumor nodule(s) in a contralateral lobe |
| | M1b | Distant Metastasis (in extrathoracic organs) |

Table1: Tumor-Node-Metastasis (TNM) Staging System

[5] Based on author's publication no. 1 and no. 9 [Appendix D]

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Stage 0 | $T_{is}$ | N0 | M0 | Stage IIB | T2b | N1 | M0 | Stage IIIB | T1a | N3 | M0 |
| Stage IA | T1a | N0 | M0 | | T3 | N0 | M0 | | T1b | N3 | M0 |
| | T1b | N0 | M0 | Stage IIIA | T1a | N2 | M0 | | T2a | N3 | M0 |
| Stage IB | T2a | N0 | M0 | | T1b | N2 | M0 | | T2b | N3 | M0 |
| Stage IIA | T2b | N0 | M0 | | T2a | N2 | M0 | | T3 | N3 | M0 |
| | T1a | N1 | M0 | | T2b | N2 | M0 | | T4 | N2 | M0 |
| | T1b | N1 | M0 | | T3 | N1 | M0 | | T4 | N3 | M0 |
| | T2a | N1 | M0 | | T3 | N2 | M0 | Stage IV | Any T | Any N | M1a |
| | | | | | T4 | N0 | M0 | | Any T | Any N | M1b |
| | | | | | T4 | N1 | M0 | | | | |

Table2: AJCC 7th Edition Staging System of Lung Cancer

## 7.1 Methodology

Extracted features from the image collection have been stored in a Comma Separated Value (CSV) file. In this way, the data file becomes platform-independent and can be accessed by any known data analysis or statistical packages. The clinical information from the original dataset has been extracted (Figure 39) and clubbed with the extracted features as per case-identification numbers. Four different datasets have been prepared: the T-staging information, M-staging information, N-staging information, and AJCC staging (Figure 40) information are appended in the feature set as the class labels, respectively. Another file has been prepared for inserting histopathological grading information as the class or target variable with the extracted features as per the concerned case identifier. Thus, five different data files have been prepared. All missing values for nominal and numeric attributes in the dataset are replaced by the modes (for nominal data) and means (for numeric data), respectively.

| Case ID | Gender | Histology | Pathological T stage | Pathological N stage | Pathological M stage | AJCC Staging (Version 7) | Histopathological Grade | Chemotherapy |
|---|---|---|---|---|---|---|---|---|
| R01-021 | Female | Adenocarcinoma | T1a | N0 | M0 | IA | G2 Moderately differentiated | No |
| R01-022 | Male | Adenocarcinoma | Tis | N0 | M0 | IA | G2 Moderately differentiated | No |
| R01-023 | Male | Adenocarcinoma | Tis | N0 | M0 | IA | G2 Moderately differentiated | No |
| R01-024 | Male | Adenocarcinoma | T2a | N0 | M0 | IIA | G1 Well differentiated | No |
| R01-025 | Male | Adenocarcinoma | T2a | N0 | M0 | IIA | G1 Well differentiated | No |
| R01-026 | Male | Adenocarcinoma | T1b | N0 | M0 | IA | G2 Moderately differentiated | Yes |
| R01-027 | Male | NSCLC NOS (not otherwise | T4 | N0 | M0 | IIIA | G3 Poorly differentiated | No |
| R01-028 | Male | Adenocarcinoma | T2a | N0 | M0 | IB | G3 Poorly differentiated | No |
| R01-029 | Male | NSCLC NOS (not otherwise | T3 | N2 | M0 | IIIA | G2 Moderately differentiated | Yes |
| R01-030 | Male | Adenocarcinoma | T1a | N2 | M0 | IIIA | G1 Well differentiated | No |
| R01-031 | Male | NSCLC NOS (not otherwise | T1b | N0 | M0 | IA | G3 Poorly differentiated | No |
| R01-032 | Male | Adenocarcinoma | T2a | N0 | M0 | IB | Other, Type II: Moderately to poorly differentiate | No |
| R01-033 | Male | Adenocarcinoma | T1b | N2 | M0 | IIIA | Other, Type I: Well to moderately differentiated | No |
| R01-034 | Male | Adenocarcinoma | T2a | N1 | M0 | IIA | G2 Moderately differentiated | Yes |
| R01-035 | Male | Adenocarcinoma | T2a | N2 | M0 | IIIA | G1 Well differentiated | Yes |
| R01-036 | Male | Adenocarcinoma | T3 | N2 | M0 | IIIA | G3 Poorly differentiated | Yes |
| R01-037 | Male | Adenocarcinoma | T2a | N0 | M0 | IB | Other, Type II: Moderately to poorly differentiate | No |
| R01-038 | Male | Squamous cell carcinoma | T2a | N0 | M0 | IB | G2 Moderately differentiated | No |
| R01-039 | Male | Squamous cell carcinoma | T1b | N0 | M0 | IA | G1 Well differentiated | No |
| R01-040 | Male | Squamous cell carcinoma | T2a | N0 | M0 | IB | G2 Moderately differentiated | No |
| R01-041 | Female | Squamous cell carcinoma | T1a | N0 | M0 | IA | G2 Moderately differentiated | No |
| R01-042 | Male | Squamous cell carcinoma | T2b | N0 | M0 | IIA | Other, Type I: Well to moderately differentiated | No |
| R01-043 | Male | Adenocarcinoma | T1a | N0 | M0 | IA | G2 Moderately differentiated | No |
| R01-044 | Male | Adenocarcinoma | T1a | N0 | M0 | IA | G2 Moderately differentiated | No |

Figure 39: Glimpse of clinical information retrieved from NSCLC Radiogenomics

| Case ID | f1 | f2 | f3 | f4 | f5 | f6 | f7 | f8 | ... | f60 | f61 | f62 | f63 | f64 | AJCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| R01-001 | 0.000876 | 0.000452 | 0.001856 | 0.001625 | 0.001185 | 0.015315 | 0.011849 | 0.026039 | ... | 0.02235 | -0.00078 | -1.3E-05 | 0.002443 | 0.001339 | IA |
| R01-002 | 8.78E-05 | -0.00058 | 0.000641 | 0.00085 | 0.022209 | 0.021155 | 0.022783 | 0.021419 | ... | 0.024818 | 0.005619 | -0.00434 | 0.005647 | 0.00699 | IA |
| R01-003 | 0.000747 | -0.00043 | 0.000759 | 0.00043 | 0.005447 | 0.013704 | 0.006771 | 0.013865 | ... | 0.015372 | -0.00511 | -0.00192 | 0.005113 | 0.001938 | IIB |
| R01-004 | 0.000167 | 0.000159 | 0.000236 | 0.000233 | 0.032542 | 0.072547 | 0.037084 | 0.072559 | ... | 0.054769 | -9.4E-05 | -0.00028 | 0.000155 | 0.000281 | IIIA |
| R01-005 | 0.002829 | 0.001566 | 0.002927 | 0.001585 | 0.002265 | 0.008838 | 0.01169 | 0.013518 | ... | 0.004772 | -0.00035 | 0.000462 | 0.001113 | 0.000896 | IB |
| R01-006 | 6.67E-05 | 6.67E-05 | 6.67E-05 | 6.67E-05 | 0.007402 | 0.049671 | 0.009745 | 0.049742 | ... | 0.03673 | -0.00023 | -0.00012 | 0.000234 | 0.000116 | IA |
| R01-007 | 0.000344 | 0.000462 | 0.000344 | 0.000462 | 0.000964 | 0.014738 | 0.003885 | 0.014783 | ... | 0.039321 | -0.00034 | -0.00035 | 0.000376 | 0.000361 | IIA |
| R01-008 | 7.72E-05 | 3.19E-05 | 8.12E-05 | 3.69E-05 | -0.00017 | 0.00228 | 0.0008 | 0.0023 | ... | 0.010192 | -3.8E-05 | -0.00018 | 5.94E-05 | 0.000177 | O |
| R01-009 | 5.9E-05 | -0.00011 | 0.000123 | 0.000125 | 0.014552 | 0.015899 | 0.014937 | 0.016272 | ... | 0.046813 | -0.0027 | -0.00203 | 0.00577 | 0.004026 | IA |
| R01-010 | 0.00018 | 0.000497 | 0.000215 | 0.000507 | 0.039812 | 0.043307 | 0.045043 | 0.043339 | ... | 0.086461 | -0.00102 | -0.00147 | 0.001026 | 0.001467 | IIB |
| R01-011 | -0.00054 | -0.00053 | 0.002351 | 0.001612 | 0.026894 | 0.038513 | 0.031001 | 0.048141 | ... | 0.029896 | 0.000569 | 0.000924 | 0.001256 | 0.002247 | IIIA |
| R01-012 | 0.000833 | 0.002142 | 0.004797 | 0.003137 | 0.019833 | 0.029321 | 0.021403 | 0.032609 | ... | 0.055337 | -0.00304 | -0.00557 | 0.003036 | 0.005572 | IB |
| R01-013 | 0.000216 | 0.000467 | 0.000669 | 0.000778 | 0.029315 | 0.035761 | 0.029807 | 0.037592 | ... | 0.026484 | -0.00264 | -0.00242 | 0.002993 | 0.002909 | IB |
| R01-014 | 4.99E-05 | 4.99E-05 | 4.99E-05 | 4.99E-05 | 0.037701 | 0.033824 | 0.056525 | 0.033824 | ... | 0.03148 | -5E-05 | -5E-05 | 4.99E-05 | 4.99E-05 | O |

Figure 40: AJCC staging information clubbed as the class variable with the extracted feature set as per case ID

## 7.2 Standardization and Normalization

Feature standardization implies that each attribute in the data set will have zero-mean and a unit variance:

$$x' = \frac{x - \bar{x}}{\sigma} \quad \text{…Equation 7.2.1}$$

Where x is the original feature vector, $\bar{x}$ is the mean of that feature vector, $\sigma$ is its standard deviation and $x'$ is the standardized value. The min-max method of rescaling was adopted to confine the values of features within the range [0, 1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad \text{…Equation 7.2.2}$$

Where x is an input value of a feature and $x'$ is the normalized value of the same. First, the dataset was separately standardized (Equation 7.2.1) (Figure 41) to the range [-1, 1] and then normalized (Equation 7.2.2) (Figure 42) to the range [0, 1].

| Case ID | f1 | f2 | f3 | f4 | f5 | f6 | f7 | f8 | ... | f60 | f61 | f62 | f63 | f64 | AJCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| R01-001 | 0.578774 | 0.341802 | 0.907862 | 0.946885 | -0.61278 | -0.1207 | -0.07568 | 0.224436 | ... | -0.00885 | -0.15262 | 0.383916 | 0.540587 | -0.03929 | IA |
| R01-002 | -0.10001 | -0.63247 | -0.05705 | 0.209487 | 1.134052 | 0.136475 | 0.704956 | 0.026082 | ... | 0.104023 | 2.896317 | -1.85726 | 2.003236 | 2.656619 | IA |
| R01-003 | 0.467441 | -0.48853 | 0.03694 | -0.1898 | -0.25864 | -0.19164 | -0.43828 | -0.29827 | ... | -0.328 | -2.21354 | -0.60477 | 1.759495 | 0.24642 | IIB |
| R01-004 | -0.03201 | 0.065766 | -0.37844 | -0.37765 | 1.99261 | 2.399467 | 1.726054 | 2.221893 | ... | 1.473962 | 0.176439 | 0.245499 | -0.50404 | -0.54412 | IIIA |
| R01-005 | 2.260805 | 1.390013 | 1.75832 | 0.90859 | -0.52302 | -0.40589 | -0.08704 | -0.31317 | ... | -0.81281 | 0.053237 | 0.630399 | -0.06653 | -0.25068 | IB |
| R01-006 | -0.11817 | -0.02085 | -0.51321 | -0.53583 | -0.09623 | 1.392135 | -0.22594 | 1.242198 | ... | 0.648863 | 0.109633 | 0.330818 | -0.46786 | -0.62266 | IA |
| R01-007 | 0.120809 | 0.351224 | -0.29281 | -0.15961 | -0.63115 | -0.14608 | -0.6443 | -0.25884 | ... | 0.767368 | 0.058889 | 0.207516 | -0.4031 | -0.50577 | IIA |
| R01-008 | -0.10911 | -0.05356 | -0.50164 | -0.56412 | -0.72499 | -0.69468 | -0.86456 | -0.79484 | ... | -0.56491 | 0.202851 | 0.299351 | -0.54757 | -0.5937 | O |
| R01-009 | -0.12479 | -0.18875 | -0.46834 | -0.48078 | 0.49783 | -0.09496 | 0.144751 | -0.19492 | ... | 1.110072 | -1.06435 | -0.66285 | 2.059356 | 1.242724 | IA |
| R01-010 | -0.02093 | 0.384362 | -0.39535 | -0.11667 | 2.596569 | 1.111917 | 2.294274 | 0.967282 | ... | 2.923516 | -0.26611 | -0.36931 | -0.1064 | 0.021816 | IIB |
| R01-011 | -0.64425 | -0.57761 | 1.301347 | 0.934338 | 1.523284 | 0.900813 | 1.291683 | 1.173476 | ... | 0.336312 | 0.491821 | 0.86973 | -0.0012 | 0.393991 | IIIA |
| R01-012 | 0.541997 | 1.932077 | 3.243887 | 2.385588 | 0.936639 | 0.496033 | 0.606426 | 0.506562 | ... | 1.49994 | -1.22466 | -2.49695 | 0.811402 | 1.980349 | IB |
| R01-013 | 0.010111 | 0.355443 | -0.03487 | 0.140942 | 1.72442 | 0.77964 | 1.206458 | 0.720517 | ... | 0.180229 | -1.03382 | -0.86504 | 0.791494 | 0.709644 | IB |
| R01-014 | -0.13257 | -0.03658 | -0.52649 | -0.55173 | 2.421225 | 0.694337 | 3.114052 | 0.558719 | ... | 0.408743 | 0.19728 | 0.365006 | -0.55189 | -0.65413 | O |

Figure 41: Standardized Feature set clubbed with AJCC staging

| Case ID | f1 | f2 | f3 | f4 | f5 | f6 | f7 | f8 | ... | f60 | f61 | f62 | f63 | f64 | AJCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| R01-001 | 0.496845 | 0.582399 | 0.22489 | 0.248781 | 0.114322 | 0.286713 | 0.174441 | 0.228395 | ... | 0.198125 | 0.400864 | 0.560865 | 0.151622 | 0.118513 | IA |
| R01-002 | 0.442893 | 0.496058 | 0.077675 | 0.13013 | 0.472189 | 0.329647 | 0.338196 | 0.187094 | ... | 0.220002 | 0.646873 | 0.334334 | 0.35047 | 0.618772 | IA |
| R01-003 | 0.487996 | 0.508815 | 0.092014 | 0.065883 | 0.186874 | 0.274871 | 0.098377 | 0.119559 | ... | 0.13627 | 0.234574 | 0.460931 | 0.317333 | 0.17153 | IIB |
| R01-004 | 0.448298 | 0.557937 | 0.028641 | 0.035658 | 0.648078 | 0.707437 | 0.552394 | 0.644301 | ... | 0.485515 | 0.427415 | 0.546874 | 0.009605 | 0.024835 | IIIA |
| R01-005 | 0.63054 | 0.675293 | 0.354644 | 0.242619 | 0.132711 | 0.239103 | 0.172056 | 0.116455 | ... | 0.042307 | 0.417474 | 0.585778 | 0.069085 | 0.079286 | IB |
| R01-006 | 0.441449 | 0.55026 | 0.008078 | 0.010206 | 0.220147 | 0.53927 | 0.14292 | 0.440311 | ... | 0.3256 | 0.422024 | 0.555498 | 0.014524 | 0.010261 | IA |
| R01-007 | 0.460444 | 0.583234 | 0.041705 | 0.070741 | 0.110559 | 0.282475 | 0.05516 | 0.127768 | ... | 0.348568 | 0.41793 | 0.543035 | 0.023327 | 0.031953 | IIA |
| R01-008 | 0.44217 | 0.547362 | 0.009845 | 0.005653 | 0.091334 | 0.190891 | 0.008955 | 0.016163 | ... | 0.090353 | 0.429546 | 0.552317 | 0.003687 | 0.015636 | O |
| R01-009 | 0.440923 | 0.535381 | 0.014925 | 0.019063 | 0.341849 | 0.29101 | 0.220681 | 0.141078 | ... | 0.414988 | 0.327299 | 0.455061 | 0.3581 | 0.356407 | IA |
| R01-010 | 0.449179 | 0.586171 | 0.02606 | 0.07765 | 0.771808 | 0.49249 | 0.671591 | 0.383069 | ... | 0.766459 | 0.391707 | 0.484731 | 0.063664 | 0.129852 | IIB |
| R01-011 | 0.399634 | 0.50092 | 0.284924 | 0.246762 | 0.551929 | 0.457247 | 0.461275 | 0.426002 | ... | 0.265023 | 0.452862 | 0.609969 | 0.077966 | 0.198914 | IIIA |
| R01-012 | 0.493922 | 0.723331 | 0.581295 | 0.480274 | 0.431745 | 0.389672 | 0.317527 | 0.287138 | ... | 0.49055 | 0.314364 | 0.269676 | 0.18844 | 0.493282 | IB |
| R01-013 | 0.451646 | 0.583608 | 0.081059 | 0.119101 | 0.593135 | 0.437018 | 0.443397 | 0.331688 | ... | 0.234772 | 0.329763 | 0.434624 | 0.185733 | 0.257487 | IB |
| R01-014 | 0.440305 | 0.548867 | 0.006053 | 0.007647 | 0.735886 | 0.422778 | 0.843557 | 0.297999 | ... | 0.279061 | 0.429096 | 0.558953 | 0.0031 | 0.004422 | O |

Figure 42: Normalized feature set clubbed with AJCC staging

## 7.3 Data Resampling

The dataset has been resampled to avoid the effect of sampling bias and class imbalance. Synthetic Minority Oversampling TEchnique (SMOTE) [142] was used as the data resampling technique in the study. If a dataset has s samples and f features in the feature space, to oversample, a sample is taken from the dataset, and its k nearest neighbors are considered. A vector between one of the k neighbors is taken. The vector is multiplied by a random number x which lies between 0, and 1. This is added to the current data point to create the new, synthetic data point. The SMOTE algorithm is as follows:

Input: no. of minority class samples (s); % of SMOTE (p); no. of nearest neighbors (k); no. of attributes (t)

Output: (p/100)*s synthetic minority class samples

*Step 1: If p<100 then randomize the minority class samples p*

*Step 2: the amount of SMOTE is assumed to be a multiple of 100*

*Step 3: initialize an index to 0. This will count the no. of synthetic samples generated*

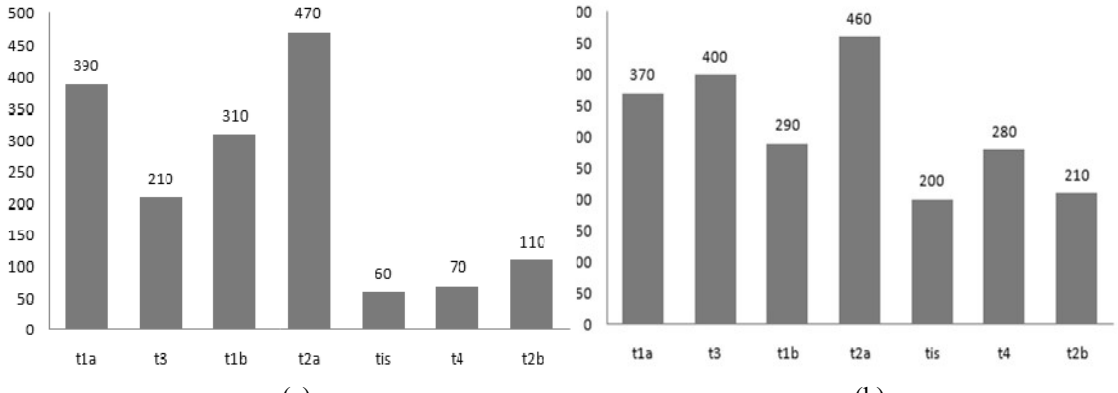
*Step 4: initialize an array for storing synthetic samples*

*Step 5: for each minority class samples compute k, save the indices and populate the synthetic samples array for each t*

At each resampling iteration, the index of the non-empty minority class value (to which SMOTE should be applied) has been selected along with 5 as the number of nearest neighbors and the average percentage of SMOTE instances to be created is 100 (with a random seed value of 1). In this way, the final dataset is prepared and is ready to be fed into any machine learning process. For example, T-stage information is clubbed with the feature set as the class variable (Figure 43). Such resampled data may be fed into any machine learning algorithm and the outcome obtained would be less biased. Thus, the result will be more robust and trustworthy, which may be compared with other machine learning models. In this way, the performance of the newly developed model becomes more reliable.

| f1 | f2 | f3 | f4 | f5 | ... | f25 | f26 | f27 | f28 | f29 | f30 | f31 | f32 | f33 | T-Stage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.006897 | 0.082764 | 0.059435 | 79 | 67.70797 | ... | 0.17429 | 0.068662 | 0.090976 | 0.149021 | 0.166608 | 0.237764 | 0.127928 | 0.122698 | 0.099106 | T1a |
| 0.010132 | 0.100149 | 0.081666 | 73.44578 | 49.63855 | ... | 0.190752 | 0.29852 | 0.10056 | 0.062611 | 0.032788 | 0.014766 | 0.024693 | 0.120736 | 0.29852 | T1a |
| 0.023682 | 0.15206 | 0.161641 | 65.54897 | 64.35567 | ... | 0.311401 | 0.161006 | 0.062116 | 0.050244 | 0.04441 | 0.028175 | 0.045779 | 0.124013 | 0.257344 | T3 |
| 0.004578 | 0.067505 | 0.042163 | 68.76 | 73.88 | ... | 0.308194 | 0.308194 | 0.188269 | 0.141006 | 0.065096 | 0.007018 | 0.002479 | 0.005268 | 0.015152 | T1b |
| 0.018066 | 0.133196 | 0.130442 | 62.19257 | 57.47297 | ... | 0.246605 | 0.216608 | 0.141492 | 0.130275 | 0.200024 | 0.154105 | 0.074097 | 0.038471 | 0.012594 | T2a |
| 0.023193 | 0.150522 | 0.159013 | 68.28421 | 80.12105 | ... | 0.083895 | 0.078524 | 0.042517 | 0.004371 | 0.032514 | 0.323828 | 0.323828 | 0.149492 | 0.059702 | T1b |
| 0.00885 | 0.093661 | 0.07307 | 65.57931 | 64.36552 | ... | 0.207944 | 0.263077 | 0.21714 | 0.136991 | 0.00533 | 0.005793 | 0.024564 | 0.067887 | 0.105772 | T1a |
| 0.020996 | 0.143375 | 0.146997 | 78.29942 | 57.19477 | ... | 0.276753 | 0.146699 | 0.071113 | 0.095483 | 0.039818 | 0.094928 | 0.050805 | 0.088894 | 0.152309 | Tis |
| 0.016235 | 0.126383 | 0.119746 | 62.38722 | 74.08271 | ... | 0.006598 | 0.029807 | 0.018356 | 0.031114 | 0.033875 | 0.107816 | 0.1863 | 0.265212 | 0.265212 | T1a |
| 0.010498 | 0.101924 | 0.084077 | 73 | 56.53488 | ... | 0.004886 | 0.117577 | 0.051444 | 0.054425 | 0.229287 | 0.295642 | 0.284459 | 0.224459 | 0.064874 | T3 |
| 0.017212 | 0.130064 | 0.125486 | 82.93617 | 72.71986 | ... | 0.134222 | 0.143584 | 0.192036 | 0.160101 | 0.212792 | 0.092397 | 0.09633 | 0.099099 | 0.125454 | T1a |
| 0.003418 | 0.058365 | 0.032925 | 71.5 | 86.41071 | ... | 0.047549 | 0.216612 | 0.204857 | 0.20198 | 0.094418 | 0.107752 | 0.156852 | 0.25567 | 0.25567 | T2a |
| 0.010254 | 0.100744 | 0.082472 | 74.42857 | 60.93452 | ... | 0.1631 | 0.183474 | 0.150857 | 0.141486 | 0.221363 | 0.130722 | 0.090742 | 0.092472 | 0.118165 | T2a |

Figure 43: A glimpse of the feature set used in the study with T-Staging as the class variable



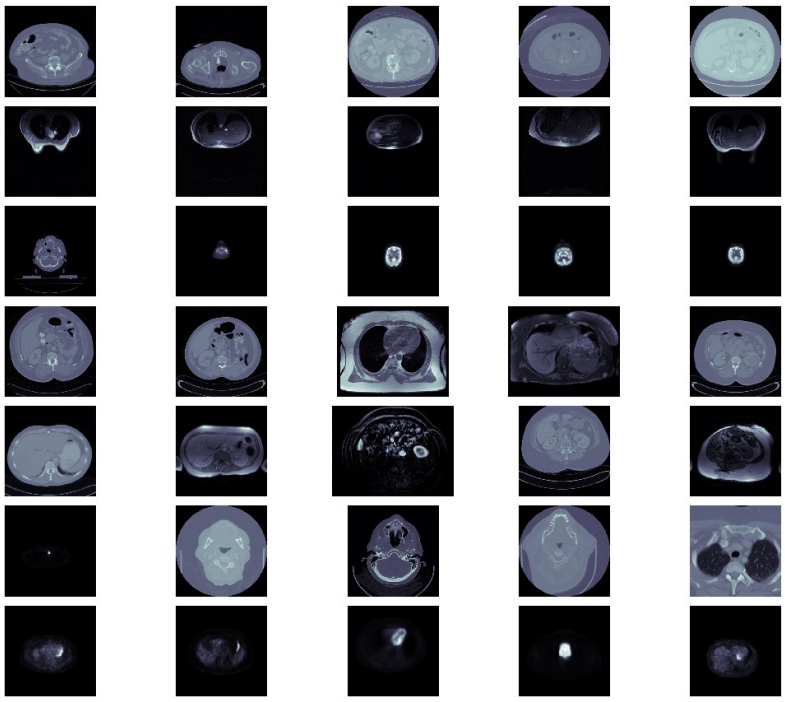

(a) (b)

Figure 44: Sample distribution under T-stage class (a) before re-sampling, (b) after re-sampling by SMOTE

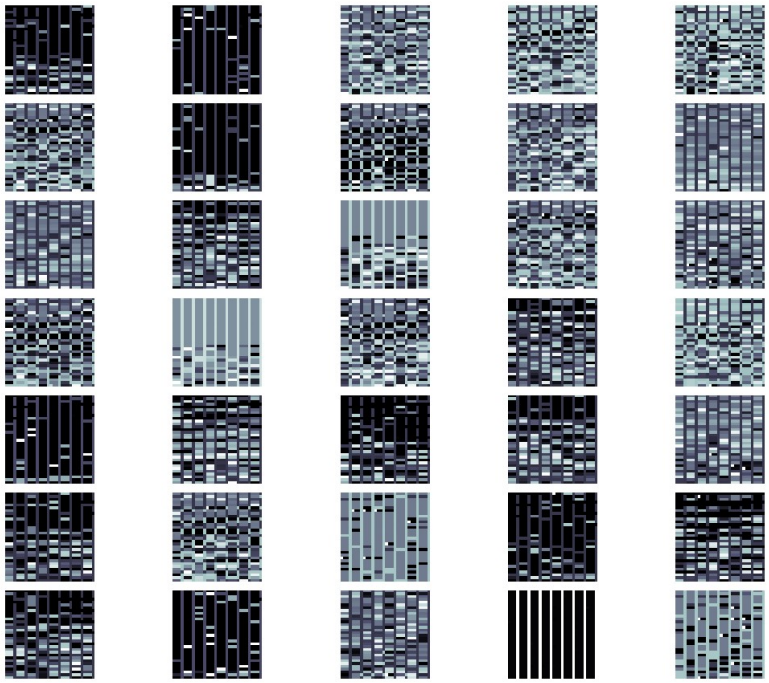
From Figure 44, it may be observed that after applying SMOTE, the class imbalance scenario has reduced to a larger extent. This has also reduced the chance of overfitting and underfitting while carrying out experiments with different machine learning algorithms.

## 7.4 Imagery Database Preparation

The main course of database preparation that has been discussed so far in the study is meant for use with traditional machine learning methods and a one-dimensional Convolutional Neural Network (1-D CNN). Such a database preparation follows manual feature extraction and gives birth to a semi-automated tumor classification system. This type of system is very useful where hardware resources are limited. On the other hand, if there is no hardware constraint, a fully automated system may be implemented by using a two-dimensional convolutional neural network (2-D CNN) or higher-level architectures as applicable. Database preparation for such a system is somewhat different and involves a specific flow of activities. Here, images are being used directly rather than extracting features into a repository. Images are labeled according to their CaseIDs and stored in a particular folder. Class or target variables of each of the images are stored in a CSV file as per the CaseID. Thus, there is a scope of mapping between the images and their respective class variables, e.g., AJCC staging. First images are read and converted into a pixel array. Later, these pixel arrays are downsampled (Figure 45), reshaped, and again transformed into image arrays. After that, image arrays are tagged with the corresponding class value and compressed for future use (Figure 46). Compression reduces the memory overhead while carrying out experiments.



(a)



(b)

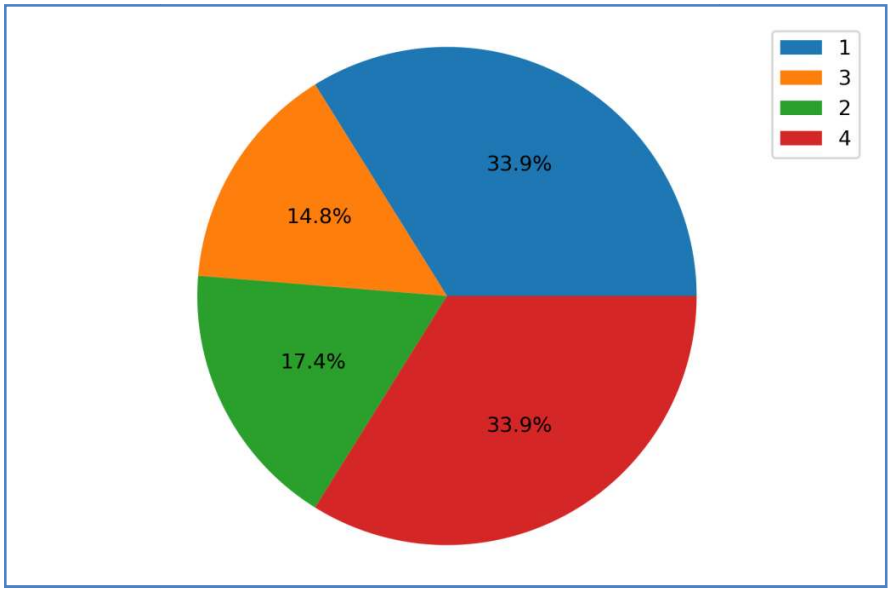
Figure 45: Image transformation during down-sampling phase at a glance - (a) original image collection (b) transformed image collection having distinct patterns for each class of tumors

In the case of image regeneration, images are labeled and directly fed into the machine learning model after resizing and re-sampling. As the outcome, regenerated images with calculated labels may be observed. Such a system works like an encoder-decoder system with an unsupervised learning mode. In the present research work, labels are tagged with the imagery array and fed into different machine learning models. Models work like encoders with supervised learning mode. The outcome may be decoded by using an inverse transformation to generate the stage and grade of a tumor.

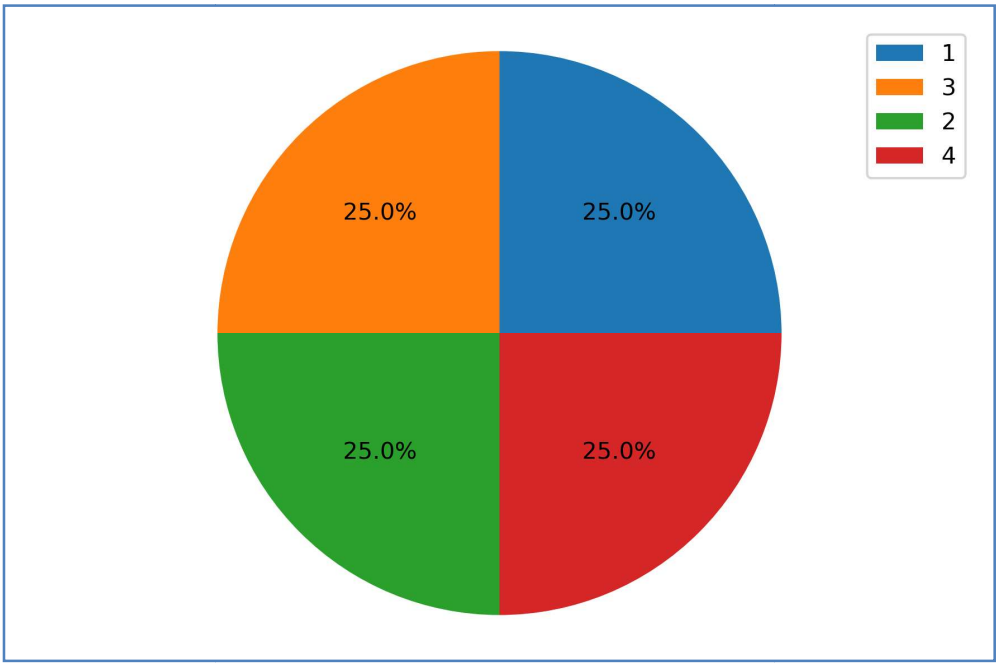
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Image Array | [[[[ 0.0000e+00]<br>[ 0.0000e+00]<br>[ 0.0000e+00]<br>...<br>[ 4.9176e+04]<br>[ 4.9176e+04]<br>[ 1.6184e+04]] | [[ 2.4000e+01]<br>[ 2.4000e+01]<br>[ 2.4000e+01]<br>...<br>[ 2.3000e+01]<br>[ 2.3000e+01]<br>[ 1.6183e+04]] | [[ 1.6424e+04]<br>[ 1.6407e+04]<br>[ 1.6407e+04]<br>...<br>[ 2.4000e+01]<br>[ 2.4000e+01]<br>[ 1.6184e+04]] | ... | [[ 2.9725e+04]<br>[ 2.9725e+04]<br>[ 2.9725e+04]<br>...<br>[ 2.2045e+04]<br>[ 2.2045e+04]<br>[ 1.6285e+04]] | [[ 2.3581e+04]<br>[ 2.3581e+04]<br>[ 2.3581e+04]<br>...<br>[ 2.8701e+04]<br>[ 2.8701e+04]<br>[ 1.6285e+04]] | [[ 2.5629e+04]<br>[ 2.5629e+04]<br>[ 2.5629e+04]<br>...<br>[ 1.4879e+04]<br>[ 1.4879e+04]<br>[-1.6465e+04]]]] |
| AJCC Staging | 1 | 1 | 3 | | 1 | 1 | 2 |

Figure 46: An instance of image arrays and their corresponding class labels (AJCC Staging)

Similar datasets may be developed by using the same image collection and by altering class variables as per need. At last, different varieties of Synthetic Minority Over-sampling Technique (SMOTE) [143] have been applied to balance class distribution.




(a)




(b)

Figure 47: Ratio of samples under different class (a) before and (b) after resampling [1=> AJCC Stage I; 2=> AJCC Stage II; 3=> AJCC Stage III; 4=> AJCC Stage IV]

First, the oversampling has been done with the minority classes and then resampling has been performed with borderline classes. At last, resampling has been done by using all the classes including both majority and minority classes. This has balanced the class distribution and helped in generating a consistent result.

## 7.5 Conclusion

This chapter describes the database preparation from features extracted in the earlier phase of the study. Vector or sequence processing is an area where deep learning is a bit underexplored than the traditional machine learning methods. The present research work has prepared a manual dataset of rank one tensor which may be fed into a one-dimensional CNN and other leading machine learning techniques. This has been done by fetching pathological staging or grading information from the clinical data and tagging those with the respective feature records. For fully automated DNN, the database has been prepared by mapping clinical information with the image array. Separate databases have been prepared for TNM staging, AJCC staging, histopathological grading, and histological subtypes.