Information loss is a common issue in Privacy-Preserving Big Data Publishing. With current privacy and anonymization techniques, data can be shared for public use and analysis without compromising individual privacy. However, this is usually at the cost of considerable information loss, where future analysis loses out on possible insights. The method proposed in this paper, Enhanced Stratified Sampling (ESS), is a solution to this common tradeoff between privacy and information loss. ESS preserves privacy while sustaining information and data utility. Additionally, it is parameter-free, identifying the best values for *k*-anonymity and *l*-diversity within the method instead of requiring them to be specified beforehand, which could otherwise introduce bias or error to the process. ESS also offers efficiency, with the method running in polynomial time, depending exclusively on the size of the dataset and the variety of unique sensitive attribute values. Furthermore, the space complexity only depends on the set's dimensions.

ESS aims to improve upon preexisting data privacy techniques by combining and automating methods to turn data into usable equivalence classes (ECs) without user input. This process begins with any necessary data cleaning that removes repeated instances, errors or inconsistencies, and explicit identifiers. The dataset should be made up of records formatted as tuples with quasi-identifiers, sensitive attributes (SAs), and any other attributes that are included with the data. Once the data is cleaned, it gets stratified based on the number of unique sensitive attribute values in the dataset. In other words, the data gets broken up into groups that share the same sensitive attribute value. Once the data is stratified, equivalence classes are created. The number of equivalence classes generated is based on the length of the original data divided by the k-anonymity level, which also serves as the minimum size of an EC. Additionally, each group is then sorted on quasi-identifier values, putting the records with the most similar values as neighbors. The algorithm then selects proportional amounts of data from each group to be placed into each equivalence class. This is ESS' method of achieving t-closeness. By taking equal portions from each data group, the distribution of sensitive attribute values within an equivalence class will be less than or equal to the distribution of that value within the overall data. Furthermore, once each equivalence class is populated, privacy and information loss will be calculated and stored to observe how much is lost when the data is grouped into equivalence classes. This process happens for all 'k' values between two and 'l,' the number of unique sensitive attribute values (inclusive). As privacy and information loss metrics are stored after each population, the algorithm can determine which 'k' value mitigates loss in the data. Combined loss is utilized, where privacy and information loss are given equal weights and multiplied by each other. This determines the 'k' value at which the tradeoff between privacy and data utility is most efficiently mitigated within the set of equivalence classes. At this point, the sampling algorithm is run once more at the chosen 'k' value, and the returned set of equivalence classes will be the one used in publishing. The algorithm's time complexity is O(nl) as it only depends on the size of the dataset (n) and the diversity of the sensitive attribute column (l); space complexity is O(n*m) as it only depends on the dimensionality of the dataset (m).

In summary, Enhanced Stratified Sampling optimizes equivalence class generation by automating the process of determining the best ‘k’ value to create classes on, without any specific privacy parameters guiding it. Its performance is only dependent on the dimensionality of the dataset; it balances the tradeoffs between k-anonymity, l-diversity, and t-closeness methods, leading to better privacy but also more data utility. This algorithm works well with one sensitive attribute, but future research needs to be performed to examine how it can be applied to datasets with two sensitive attributes, which can then be extended to any number of sensitive attribute columns.